Sex Differences in Cognitive Abilities and Educational Outcomes: Examining the Contribution of Sex-Role Identification

David Hugh Reilly
BPsych (Hons) BSoft. Eng. (Hons)

School of Applied Psychology
Griffith Health
Griffith University, Gold Coast

Submitted in fulfilment of the requirements of the degree of
Doctor of Philosophy in Applied Psychology
February 2019
Abstract

Sex differences in cognitive ability have been documented in psychological research for over a century, and the research area has seen considerable changes in theoretical perspectives and methodology. While males and females do not differ in general intelligence, an extensive body of literature documents sex differences in more specific cognitive tasks (for reviews see Halpern, 2000; Kimura, 2000; Maccoby & Jacklin, 1974). Males on average perform at a higher level on tasks that rely on visual-spatial ability, and this has been linked to later gender gaps in quantitative abilities such as mathematics and science and to the underrepresentation of women in science, technology, engineering and mathematics (STEM)-related fields. Females as a group do better at tasks involving verbal and language abilities which have been linked to wide gender gaps in reading and writing, as well as the underrepresentation of men in post-secondary education. Some researchers have argued that sex differences in cognitive ability are declining in response to social changes in the roles and status of women, but methodological limitations and use of convenience samples have limited previous enquiries seeking to test that hypothesis.

The aim of this course of research was twofold. Firstly, using the statistical technique of meta-analysis to examine the evidence for sex differences in visual-spatial, verbal and quantitative abilities, and - given the passage of time - whether they were declining in response to changes in the roles of men and women in society. This was addressed through a series of studies that examined: i) nationally representative samples of student testing data from the National Assessment of Educational Progress (NAEP) in the United States, ii) cross-cultural samples of student testing data from the Programme for International Student Assessment (PISA). Secondly, to determine the contribution of sex-typed personality traits and behaviours (collectively referred to as sex-role identity) to the development of individual differences in visual-spatial and verbal ability.

This goal was addressed through a sequence of three experimental studies. Empirical study 1 sought to provide the most comprehensive assessment of the sex-role mediation hypothesis conducted to-date, by examining performance across a range of visual-spatial and verbal ability tasks. Subjects high in masculinity performed better on visual-spatial tasks, while subjects high in femininity performed better on verbal
language tasks. Mediation analysis showed that sex-role identification acted as a mediator of the sex difference in cognitive tasks.

Having found evidence for sex-role differences, Empirical Study 2 sought to test whether the observed sex-role differences reflected latent ability, or alternately the role of stereotype threat and task labelling on performance. The way in which a person appraises the testing situation (and the types of skills a task may require) can work hand in hand with sex-role conformity pressures to increase or to decrease task performance.

Finally, Empirical Study 3 sought to address a limitation in the existing theoretical models for sex differences in cognitive ability, namely that males and females show different patterns of self-estimation of intellectual ability (termed the male-hubris female-humility problem). Study 3 examined the contribution of sex-role identity to self-estimated intelligence, as well as the accuracy of personal judgements of ability by administering the Cattell’s Culture Fair Test of Intelligence. Results showed that the degree of masculine identification predicted self-estimated intelligence scores.

A large body of research has shown that self-appraisal of intellectual abilities and self-efficacy beliefs guide the selection of coursework in secondary and tertiary education and form an important component of career decision-making. This may explain to some degree gender-specific differences in certain fields of STEM.

Collectively, the results of these studies are used to refine existing psychobiosocial models of sex differences in cognitive abilities, to explain both the differences between males and females but also within-sex variability. It suggests masculine and feminine sex-role identification is an important individual differences factor to consider, and that these shape intellectual self-image and self-efficacy beliefs.
STATEMENT OF ORIGINALITY

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

[signature redacted]

David H. Reilly,

December 2018
Acknowledgement of Published and Unpublished Papers Included in this Thesis

Section 9.1 of the Griffith University Code for the Responsible Conduct of Research (“Criteria for Authorship”), in accordance with Section 5 of the Australian Code for the Responsible Conduct of Research, states:

To be named as an author, a researcher must have made a substantial scholarly contribution to the creative or scholarly work that constitutes the research output, and be able to take public responsibility for at least that part of the work they contributed. Attribution of authorship depends to some extent on the discipline and publisher policies, but in all cases, authorship must be based on substantial contributions in a combination of one or more of:

- conception and design of the research project
- analysis and interpretation of research data
- drafting or making significant parts of the creative or scholarly work or critically revising it so as to contribute significantly to the final output.

Section 9.3 of the Griffith University Code (“Responsibilities of Researchers”), in accordance with Section 5 of the Australian Code, states:

Researchers are expected to:

- Offer authorship to all people, including research trainees, who meet the criteria for authorship listed above, but only those people.
- accept or decline offers of authorship promptly in writing.
- Include in the list of authors only those who have accepted authorship
- Appoint one author to be the executive author to record authorship and manage correspondence about the work with the publisher and other interested parties.
- Acknowledge all those who have contributed to the research, facilities or materials but who do not qualify as authors, such as research assistants, technical staff, and advisors on cultural or community knowledge. Obtain written consent to name individuals.

Included in this thesis is the paper in Chapter 6 for which I am the sole author. Additionally, included in this thesis are the papers in Chapters 3, 4, 5, 7 and 8 which are co-authored with other researchers. My contribution to each co-authored paper is outlined at the front of the relevant chapter, and I acknowledge the support and guidance of my supervisors in preparing these manuscripts. The bibliographic details of these publications are included for each chapter, along with a copyright statement.
List of Publications and Conference Papers Arising from this PhD Research Programme

Journal articles

(Listed in order in which these articles appear in this thesis):


Additional Articles and Conference Papers referenced (but not included due to space requirements)


List of Figures

Chapter 1: Introduction
Figure 1.1. Nash’s (1979) sex-role mediation hypothesis ...........................................10
Figure 1.2. New student enrolment in higher education diplomas and degrees in Australia, separated by gender..........................................................16

Chapter 2: Literature Review
Figure 2.1 Sex differences in mathematical performance on the SAT-M over the past two decades (1996-2016).................................................................62

Chapter 9: Empirical Study 2
Figure 9.1. Sex-role mediation theory of cognitive abilities.................................144
Figure 9.2 Mean performance on the Group Embedded Figures Test ......................155
Figure 9.3 Mean performance on the Vandenberg Mental Rotation .........................157
Figure 9.4 Mean completion time for the Vandenberg Mental Rotation Task .........158
Figure 9.5 Mean number of words generated in verbal fluency task .................160

Chapter 10: Empirical Study 3
Figure 10.1 Stimulus material used for self-estimation of intelligence ...............182
Figure 10.2 Distribution of measured IQ scores in the sample .........................187
Figure 10.3 Rosenberg General Self-Esteem scores across sex/sex-role categories ...188
Figure 10.4 Self-estimated IQ scores across sex-role categories, for males and females ..........................................................190
Figure 10.5 Scatterplot of association between self-estimated and psychometric IQ, for males and females respectively ......................................................192
Figure 10.7 Indirect effect of sex on SEI, with masculine sex-roles acting as a mediator195

Chapter 11 Discussion
Figure 11.1 Biosociocultural model of sex-role identification acquisition ...............240
Figure 11.2 Sex-role mediation theory of cognitive development ......................242
Acknowledgements

I will be eternally grateful to my primary supervisor, Professor David Neumann for his guidance, encouragement and support throughout my candidature. Not every supervisor pairing is ideal, but I sincerely believe that this one was bespoke in every way. David encouraged me to publish early, and often, which gave my work a wider scrutiny through the peer-review process and offered excellent grounding in academic publishing. His encouragement and support meant a great deal to me as a fledgling researcher: David grants his research students with independence and autonomy but at the same time is always waiting in the wings to scaffold and support them when they need it. During my initial data collection he’d often pop on by to check in, and then as he saw I grew in confidence the visits would be more spaced out, and sometimes he’d just stick his head in the door unobtrusively to see that all was proceeding well. It might seem like a small thing, but to know that your supervisor has confidence in you means a great deal. His door was always open, and one of the most wonderful things about his supervision (and talking with fellow PhD and honours students, it’s a commonly recurring theme) is that despite his extremely busy teaching, research and administrative schedule, without exception, he makes time for everyone. His notes and revisions on chapters and manuscripts are always carefully considered, showing depth of thought, and his feedback was always supportive and encouraging. And having been privileged to have had him as a statistics lecturer during my undergraduate, I have tried to emulate (with a few embellishments of my own) his teaching style. It was a truly strange experience, that transition from student/teacher to then an equal collaborator on a research project, and then an independent researcher. I am so very grateful that he took a chance on me.

Likewise I am immensely grateful to have had Assoc. Professor Glenda Andrews as my associate supervisor. To have one competent supervisor is a boon, but
to have two was a genuine blessing. At every stage of my candidature Glenda has been supportive and encouraging, and brought to bear her expertise with cognitive and educational psychology, especially in the design and planning of studies. For example, I might not have included psychometrically measured intelligence in the form of the Cattell Culture Fair IQ Test but for her, which added greater depth to the study and its conclusions. Her contributions have been invaluable, and just like David she invested a significant amount of time, thought and energies into this programme of research.

Glenda also has such a keen eye for grammar and style as well, and was a real asset in preparing manuscripts for submission. I’ve sometimes joked with her that she missed her calling as a book-editor, because even after the umpteen revision she will spot a way to tighten the writing (clearly, by the length of this acknowledgment, brevity is not my forte!) or make its meaning clearer. When suggesting changes, Glenda also took the extra time to explain why and provide really detailed feedback – I know that it has made me a better writer as a result, and for that I am extremely grateful.

I’d also like to acknowledge the contribution of my original supervision team, Professor Liz Conlon as primary supervisor and Dr Heather Green as associate supervisor. Heather’s familiarity with the literature on sex differences in cognitive ability was invaluable early on in guiding me to particular books and studies that were important. I brought to the project passion and idealism, but Liz tempered this with scientific objectivity and healthy scepticism of the research literature: identifying a number of serious gaps in the research literature (such as the significant passage of time that had occurred, and some prominent researchers who questioned whether sex differences were present in modern samples). Much of the meta-analytic reviews in this thesis were to plug those gaps in the literature that she identified, and provide objective and comprehensive tests of sex differences and similarities in cognitive abilities. So while the research was carried out independently, you can still see echoes of her
contributions there that should be acknowledged. Liz has been, and remains still, a researcher whom I greatly admire, and I am grateful for her approval.

On a personal note, I’d also like to thank my loyal family who have been with me and supporting me on this journey all the way. Anyone who has embarked on postgraduate study knows that it is very much a team effort, and there steadfast support has meant so much. It is with great sadness that not all of them will be able to see me graduate though. During my candidature my father was diagnosed with aggressive lung cancer and passed within a few months of his initial diagnosis. He was always fiercely proud of my accomplishments and I know that he will be looking on.

My mother was also diagnosed a few years earlier with a substantially sized malignant brain tumour, and I am immensely grateful for those extra years we have experienced together. I have cherished every day with her, and though most people believe me to be her carer it’s really always been the other way round. She was the best mother a son could ever wish for, and though I tell her this often I know there’s a part of her that feels unworthy of such praise. I wish but for a moment she could see herself through my eyes. She understands me better than I know myself, and is my dearest friend and confidant. I am the man I am today because of her, and it’s with no small measure of pride that I can see so very much of her personality, worldview, and intellect in myself. She was a primary school teacher for just over 50 years, and my love of educational psychology and learning comes from her.

Her bravery and stoicism is inspiring, and in the original draft of these acknowledgements I’d written that I’m so grateful that she’ll be there to see me finally graduate. But just days before completing my PhD, we received the news that there was a new tumour in her right hemisphere. We had been so happy, and now the future is so uncertain. But I made her a promise to finish this PhD, and at least that is a promise I have been able to keep to her. She brings such joy (aptly named), and I cannot conceive
of a world without her in it. It has been said that there’s no greater tragedy than the loss of a child to a parent. But those who say this did not have Roma Joy Reilly as a mother. Though it may seem selfish, I’ve long hoped that she would outlive me so that I would be spared the pain of losing her. I yet hope that this will be the case, many many years hence!

My twin brother, Michael, has always been my greatest champion and I’m truly blessed to have had his support and love while travelling this road. I know that there have been times of personal self-doubt, but he has had faith in me always. It’s hard to adequately convey how deep the bond between twin brothers can be…. it endures throughout all the decades of your life (and hopefully beyond). He’s always been the stronger of the two of us, and so caring and protective. I was blessed to have had him as my twin, and because he –hates- compliments or praise I commit to writing words he would otherwise ignore :P

I would also like to acknowledge the many researchers along the way (and my apologies if there are any I have inadvertently overlooked) who took the time to answer questions, provide the benefit of their perspectives, or provide advance access to manuscripts or research materials. Firstly the late Professor Doreen Kimura who was immensely supportive during my early review of the literature in this area and took the time to send me copies of many of her less accessible publications, copies of her 2D mental rotation test, along with an encouraging hand written letter. I also acknowledge the support of the late Professor Sandra Bem, who provided initial guidance on the use of the Bem Sex Role Inventory. I wish both these researchers were still here to read these words of gratitude.

When gathering a feel for the research landscape, I found the literature reviews and books of Professor Diane Halpern invaluable. Though I’ve not had the pleasure of meeting her in person, I found several of her lectures on the topic of sex differences on
Youtube which provided a good grounding for the ‘unanswered’ researched questions in our field. I also would like to thank Professor Alice Eagly whom I did have the pleasure of meeting in person at a conference in Amsterdam while participating in a panel discussion. Though I was only an early PhD candidate at the time (and so quite in awe meeting some of the leading experts in my field!), Professor Eagly was welcoming and treated me as a peer – something that I was immensely grateful for at the time. I’d also like to thank Professor David C. Geary kindly answered questions on evolutionary psychology and verbal abilities, as well as provided electronic copies of his book and other publications, while Professor Adrian Furnham provided copies of stimuli materials used in his experiments and answered further questions about the male hubris, female humility effect. Professor Janet Hyde also answered questions on her work, and Dr Michael Peters provided access to a redrawn version of the Vandenberg Mental Rotation Test (VMRT) stimuli material that could be electronically administered.

I’d also like to personally thank Dr Michael Borenstein for his assistance in learning the statistical technique of meta-analysis. His book, “Introduction to Meta-analysis”, was invaluable, but he also kindly answered questions about more advanced topics such as meta-regression and provided early access to a copy of a more powerful version of his incredible Comprehensive Meta-Analysis package. I also found the guides to conducting a meta-analysis by Professor Robert Rosenthal deeply useful as well. I’d also like to thank the assistance and guidance of Professor Margaret Signorella, who offered supportive and encouraging advice early on, especially about additional statistical techniques for testing publication bias effects.

One of my favourite sayings about the scientific method is by Sir Isaac Newton, is “If I have seen further, it is by standing on the shoulders of giants”, which can be traced back to the Latin phrase nanos gigantum humeris insidentes (dwarves standing on the shoulders of giants). It is an expression of humility but also a recognition and
expression of gratitude to the work of those who have gone forth before us. The phrase will be instantly recognisable to any researcher or PhD student, of course, as it has been immortalised on the front page of Google Scholar; but not everyone will know its providence or understand its true meaning. But I found myself reflecting on it often while doing literature reviews on Google Scholar, and just how much my own work has been influenced by them. The reference list in this thesis attests to their many contributions to the research literature. It has been a privilege to add – if only just a little – to that body of scientific knowledge, but it is only by the grace of those who have gone before me and laid the foundations.
# Table of contents

Chapter 1 - Introduction ................................................................................................ 1  
1.1 Definition of key terms ........................................................................................ 1  
1.1.1 Sex versus gender differences ....................................................................... 1  
1.1.2 Intelligence versus specific cognitive abilities ............................................. 3  
1.2 Background to the topic ....................................................................................... 5  
1.3 Importance of sex difference research in educational psychology ............... 11  
1.3.1 Underrepresentation of women in STEM fields, and STEM literacy .......... 11  
1.3.2 Sex differences in literacy, schooling, and entry to higher education ....... 15  
1.3.3 Summary of educational importance ........................................................... 19  
1.4 Research questions ............................................................................................ 20  
1.5.1 RQ 1: Magnitude of sex differences in cognitive abilities ........................... 22  
1.5.2 RQ 2: Contribution of sex-role identification to cognitive performance ...... 22  
1.5.3 RQ 3: Contribution of situational factors to cognitive sex differences ....... 23  
1.5.4 RQ 4: Contribution of sex-role identification to self-estimated intelligence . 23  
1.5 Overview of current research ............................................................................. 24  

Chapter 2 – Literature Review .................................................................................... 34  
2.1 Summary of Research Findings ......................................................................... 34  
2.2.1 Verbal abilities. .......................................................................................... 35  
2.2.2 Visual-spatial ability. .................................................................................. 57  
2.2.3 Quantitative ability. .................................................................................... 57  
2.2.4 Memory. ..................................................................................................... 66  
2.2 Popular Beliefs about Intelligence ..................................................................... 75  
2.1.1 Self-Estimation of Intelligence.................................................................... 76  
2.1.2 Estimation of other’s intelligence. ............................................................... 78  
2.1.3 Popular beliefs about sex differences in specific cognitive abilities .......... 79  
2.3 Theoretical perspectives on sex differences in cognitive abilities ............... 80  
2.3.1 Biological Explanations for Sex Differences ............................................... 81  
2.3.2 Psychosocial Explanations for Sex Differences ........................................... 87  
2.3.3 Macro-level Cultural Contributions ............................................................ 97  
2.3.4 Nash’s Sex-Role Mediation Theory .......................................................... 103  
2.4 Summary of Literature Review Findings ......................................................... 108  

Chapter 3 – Gender Differences in Spatial Ability ..................................................... 132  

Chapter 4 – Sex Differences in Mathematics and Science Achievement ............... 133  

Chapter 5 – Sex Differences in Reading and Writing............................................. 134  

Chapter 6 – Cross-Cultural Patterns of Reading, Mathematics and Science Literacy 135  

Chapter 7 – Meta-Analysis of Sex-Role Mediation Effect for Visual-Spatial Ability 136  
Meta-Analysis Summary and Prelude to Empirical Studies 137  

Chapter 8 – Empirical Study 1 – Sex and Sex-Role Differences in Specific Cognitive Abilities ........................................................................................................ 138  

Chapter 9 – Empirical Study 2 – Effect of Task-Labelling, Stereotype Threat, and Sex-Role Identification on Cognitive Performance ..................................................... 139  
Overview of Gender Differences in Specific Cognitive Abilities ....................... 143  
Sex-Role Mediation as an Explanation for Sex differences ............................... 144  
Sex-typing of Cognitive Tests and Task-labelling ............................................. 147  
Gender Stereotypes, and Stereotype Threat ..................................................... 147  
The Present Study ............................................................................................. 148
Method .......................................................... 150
Participants .................................................. 150
Measures ....................................................... 150
Procedure ..................................................... 153

Results .......................................................... 154
Labelling of Spatial Task .................................. 154
Stereotype Threat .......................................... 156
Stereotype Lift ............................................... 158
Manipulation Check ....................................... 160

Discussion ..................................................... 161
Sex-typing of cognitive tasks and task-labelling .... 161
Stereotype Threat/Lift .................................... 162
Implications and Limitations ......................... 164

Conclusions ................................................... 166

References .................................................... 167

Chapter 10 – Empirical Study 3 - Sex and Sex-Role Differences in Self-Estimated Intelligence (SEI) .................................................. 175
Overview ...................................................... 175

Method .......................................................... 180
Participants .................................................. 180
Procedure ..................................................... 181
Measures ....................................................... 181

Results .......................................................... 186
Sex Role Classification .................................... 186
Cattell’s Culture Fair Intelligence Test (CCFIT) .... 186
General and Academic Self-Esteem .................. 188
Self-Estimated Intelligence (SEI) Scores .......... 189
Bivariate Correlations .................................... 191
Predictors of Sex Differences in Self-Estimated Intelligence ................. 191
Self-estimates of multiple intelligences ............ 197

Discussion ..................................................... 199
Explanations for Sex Differences in Self-Estimated Intelligence .......... 200
Self-estimates of multiple intelligences ............ 203
Implications and limitations .......................... 204

Conclusions ................................................... 207

References .................................................... 209

Chapter 11 - Discussion ...................................... 216
11.1 Magnitude of sex differences in cognitive abilities .......... 217
11.2 Contribution of sex-role identification to cognitive performance .... 225
11.3 Contribution of situational factors to cognitive sex differences .... 229
11.4 Contribution of sex role identification to self-estimated intelligence 232
11.5 Collective findings and implications for theory building .......... 237
11.6 Directions for future research and limitations .................... 247
11.7 Practical Implications for Childhood Education ................. 249
11.8 Final Conclusions ...................................... 253

References .................................................... 255

Appendices ..................................................... 262
Appendix A1 – SAT Mathematics Meta-Analysis ......................... 262
Appendix A2 – SAT Verbal Meta-Analysis ................................. 263
Appendix A3 – Academic Self-Esteem Measures ......................... 264
Chapter 1 - Introduction

“The world cannot afford the loss of the talents of half its people if we are to solve the many problems which beset us” – Rosalyn Yalow, Nobel Laureate for Medicine 1977

From the beginning of formal scientific inquiry into the nature and structure of human intelligence, researchers have sought to understand how biological and environmental factors contribute to cognitive ability. It remains an important goal of cognitive and developmental psychology (Jensen, 1998; Neisser et al., 1996). One question in particular commands the attention of researchers and lay-persons alike: that of group differences between males and females in cognitive ability (Caplan and Caplan 1994; 1997; Eagly, 1995; Hyde, 2005). In particular, the past four decades have seen a rapid expansion of interest in this topic, and this has coincided with substantial changes in the status of women, advances in research methodology, and a renewed focus on sociocultural contributions to sex differences. As a result, many of the earlier assumptions about sex differences in cognitive ability have been called into question, and experimental observations of the distant past are not guaranteed to replicate in the present-day. Before moving to an overview of the topic, a brief review of nomenclature is offered for the reader.

1.1 Definition of key terms

1.1.1 Sex versus gender differences

Historically the term “human sex differences” had been used to refer to observed differences between males and females in behaviour or cognitive performance. While the term “sex differences” is still used at the present time, many researchers opt instead for the term “gender differences”. Usage of these terms varies considerably in the
literature, with some authors preferring the term *sex* to denote differences between males and females as groups (Halpern, 1994), while other authors prefer the term *gender* to denote a sociocultural origin for male-female differences (Unger, 1979). Strong and impassioned arguments have been made for both terms (Eagly, 1995; Frieze & Chrisler, 2011; Halpern, 2011), and each may be used interchangeably even by the same author, depending on editorial preference of the journal in which their work is published. When conducting a literature review, both “sex differences” and “gender differences” must be used to gain a full view of the literature, although the prevalence of their usage varies by discipline. Studies published in medical and clinical journals (for example, on the effect of sex hormones on cognition) have a preference for the term “sex differences”, while some psychological and feminist psychology journals have editorial policies specifying the term “gender differences”. In other psychology journals such as “Intelligence” the term sex differences is used almost exclusively. Kaiser (2015) has even proposed a change in nomenclature to refer to the topic as “sex/gender”, on the basis that most researchers in the field acknowledge the futility in trying to disentangle biological and psychosocial contributions, and instead embrace what Halpern (2000, 2011) has termed a psychobiosocial model of sex differences.

Following the practice of Eagly (1987), and Halpern (2000), the term “sex differences” has been employed herein without prejudice to refer to group differences between males and females. Doing so is not intended to impute that their origin is solely or even chiefly biologically determined. Previously published literature included as chapters of the thesis may also employ the term ‘gender differences’ due to the editorial policy of particular scientific journals and publishers (e.g., Frieze & Chrisler, 2011). The term “gender stereotypes” is also used to refer to commonly held stereotypes about the abilities of males and females, as these reflect sociocultural beliefs held about the
roles and abilities of men and women that are not necessarily accurate portrayals of reality.

One additional term used throughout this thesis that is important to clarify is “sex-role identification” (sometimes referred to in the literature as sex-role identity, or just sex-roles). This term is defined as the degree to which an individual incorporates stereotypically masculine or feminine personality traits, interests, and behaviours. There is wide variability in the degree to which individuals incorporate stereotypically masculine and/or feminine personality traits into their self-identity: some persons will embody traditionally masculine agentic traits, some will embody traditionally feminine expressive traits, while still others will incorporate a blend of both (termed psychological androgyny). These issues will be described in detail later, but it is important to note that variations in sex-role identification are normative in the general population and that all sex-role categories ought to be respected and privileged. It is also distinct from the diagnosis of gender identity disorder (GID; Bartlett, 2000; Zucker & Cohen-Kettenis, 2008). Some researchers (e.g., Frieze & Chrisler, 2011) prefer the terms gender-roles and gender-role identity (in the same manner as gender differences may be preferred over sex differences), but given how similar these terms are to what the DSM-V labels a pathological condition (American Psychiatric Association, 2014), the term sex-role identification has been retained to avoid confusion.

1.1.2. Intelligence versus specific cognitive abilities.

Intelligence is a complex, multi-facet construct that encompasses a broad range of cognitive abilities (Carroll, 1993; Sternberg, 2014). When men and women are compared at a population level, their level of general intelligence as measured by standardised tests of intelligence results in an equivalent mean IQ score (Halpern, 2000; Neisser et al., 1996). In one of the most comprehensive reviews of general intelligence
and assessment of mental ability, Jensen (1998) concluded that there is no evidence of sex differences in general intelligence at a population level. These findings notwithstanding, robust sex differences are found in more specific tests of cognitive abilities. For example, males as a group typically score higher than females on tests of visual-spatial reasoning, while sex differences are reversed for language-based tasks.

Some older literature has made reference to sex differences in intelligence, a practice that can lead to confusion and the mistaken impression that researchers are attempting to argue that one sex is ‘smarter’ in some way than the other (Halpern & Lamay, 2000). This can also be easily misconstrued by laypersons and media when talking about research findings. Halpern (2000, 2011) advocated the use of the term “sex differences in cognitive abilities”, as it makes clear to the reader that aptitude for specific types of tasks is being investigated rather than overall intellectual functioning. This is particularly important, given the historical legacy of scientific bias against women from the scientific community in the 19th and early 20th centuries, whereby it was claimed that women were morally and intellectual inferior to men. Such arguments were supported by flawed evidence provided by the field of phrenology, and neuroanatomical arguments over total brain size (Shields, 1975). Another well-established finding is that males typically rate their intelligence as higher than do females in self-reports (Szymanowicz & Furnham, 2011), and that women generally underestimate their intellectual achievements which Beyer (1990) has termed a self-degrading bias. For this reason, avoiding misconstrual of scientific research findings as pertaining to overall intelligence is paramount, as it might further entrench existing negative gender stereotypes about intellectual functioning. Accordingly, the term “sex differences in cognitive abilities” has been adopted herein.
1.2 Background to the topic

The issue of whether differences in cognitive ability exist between males and females, and if so how meaningful those differences are, is of profound importance to educators, parents, policy makers, and the general public who all have an interest in raising the educational standards of both boys and girls (Halpern et al., 2007a). It was once regarded that sex differences in cognitive abilities were inevitable and the result of biological differences between the brains of males and females (Kimura, 1992, 2000) – a position known as biological determinism. With further advances in neuroimaging, closer inspection reveals relatively few structural differences beyond brain volume (Aaron et al., 2005; Hines, 2011). Similarly, endogenous hormones alone explain either a relatively small amount of variance in cognitive performance (Gouchie & Kimura, 1991; Hausmann, Schoofs, Rosenthal, & Jordan, 2009) or no associations are found (Herlitz, Reuterskiöld, Lovén, Thilers, & Rehnman, 2013; Kocoska-Maras et al., 2011; Puts et al., 2010). At best, the evidence of activational effects of sex hormones is mixed and inconsistent in healthy non-clinical adults (for a review see Miller & Halpern, 2013). There is also wide variability in individual performance with some males and females performing significantly better (or significantly worse) than their same-sex peers (referred to as within-sex variability, as opposed to between-sex variability). Moreover, considerable cross-cultural variation can be seen in the magnitude of cognitive-sex differences (Guiso, Monte, Sapienza, & Zingales, 2008; Reilly, 2012; Riegle-Crumb, 2005). This variation suggests that the emergence of sex differences in the general population is at least partially determined by social and cultural processes, such as cultural beliefs and educational practices. Researchers (e.g., Hyde, 2005) have also argued that the size of sex differences has been decreasing over time in response to changes in the relative status of men and women in society. Such a premise would be
inconsistent with the notion of immutable sex differences in cognitive ability (Feingold, 1988; Hyde, Lindberg, Linn, Ellis, & Williams, 2008). Taken together, this implies that the emergence of sex differences in cognitive abilities at the population level is not fixed, and may thereby be subject to intervention.

Substantial sex differences are found for some types of cognitive tasks, but on other tasks males and females perform equally well (Zell, Krizan, & Teeter, 2015). For example, females tend to score higher on tests that measure aspects of verbal ability (Hyde & Linn, 1988), while males show higher performance on tests of spatial ability (Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995). However, sex differences in quantitative ability (Maccoby & Jacklin, 1974), such as mathematical and scientific reasoning, remain contentious. Some researchers find evidence of small but still influential differences in quantitative reasoning (Hedges & Nowell, 1995), while others argue any observed differences in maths are so small, in fact, that they can be categorised as ‘trivial’ (Hyde & Linn, 2006). Many of the inconsistencies in the literature may be the result of highly selective samples that do not generalise well to the general population (Becker & Hedges, 1988). Alternately there may be other (as yet unidentified) factors that covary across samples with sex differences in cognitive ability such as sex role identification, or socioeconomic status (Levine, Vasilyeva, Lourenco, Newcombe, & Huttenlocher, 2005; McGraw, Lubienski, & Strutchens, 2006; Nash, 1979).

The issue of whether differences in cognitive ability between males and females still exist in modern samples, and if so how meaningful are those differences, is a question of interest to multiple stakeholders (educators, parents, policy makers, and the wider society). But making such a determination requires a strong body of evidence, and in their zeal, a number of researchers have been slightly premature in declaring the
SEX AND SEX-ROLE DIFFERENCES IN COGNITIVE ABILITIES

problem solved. Feingold (1988) claimed sex differences in cognitive ability were “disappearing”, but this conclusion was not widely accepted and strongly critiqued. In an effort to clarify the matter, Hedges and Nowell (1995) published the largest analysis of student testing data for the USA ever conducted, demonstrating a pattern of sex differences across three decades that was not diminishing. More recently, Caplan and Caplan (2005, p. 25) have called investigating gender a “perseverative search for sex differences”, but in their review omitted key sources of data (such as Hedges and Nowell) that were disconfirming. Similarly, Hyde (2005) advanced the “gender similarities hypothesis” which states that psychological sex differences are either very small or trivial in magnitude. In a response, Lippa (2006) noted the selective omission of key reviews where larger effect sizes were found. For example, many of the sex differences reported in Hedges and Nowell were higher, including reading, writing, mathematics, science, and visual-spatial ability. On the subject of disappearing gender gaps, Halpern (1989, 2014) has quipped that in the investigation of sex differences, “what you see depends on where you look”.

One aim of the present programme of research is to apply the statistical technique of meta-analysis to a series of large, demographically representative datasets of student achievement that for various reasons have not been tested for sex differences, thus addressing a gap in the research literature (see meta-analysis chapters). This may help resolve some of the inconsistencies in the literature and provide greater clarity to such debates over whether the magnitudes of sex differences are decreasing over time. Educators and policy-makers need timely and well-informed evidence for decision-making. Further, resolution of these issues are a prerequisite for planning educational interventions (Liben & Coyle, 2014), as well as the allocation of limited resources by educators to address disparities in educational outcomes. But the cost of accepting a
false null hypothesis may be high – it might have a chilling effect on further research into factors contributing to sex differences in cognitive ability, as well as further development of educational interventions aimed at achieving equality of outcomes. Preiss and Hyde (2010, p. 312) warn that little research has “tested targeted interventions designed to close specific ability gaps […] though the development of such interventions are important. Regardless of the reasons for gender differences in specific cognitive abilities, whether biological, social, or cultural, it is likely that interventions can be developed to help most students”. Thus there is a lack of consensus, but this debate might be better informed with timely data.

Furthermore, when sex differences are reliably found for cognitive tasks, it poses the additional question of their origins (Halpern, 2011; Kimura, 2000; Stanley, Benbow, Brody, Dauber, & Lupkowski, 1992). While sex differences in cognitive ability have been studied since the very beginning of psychometric assessment of intelligence (for a review see Shields, 1982), theoretical perspectives on their causes have shifted considerably over this time period. Archer (1996) has termed these origin theories. Early theoretical debate on sex differences proposed strong and immutable biological factors (nature), while later theorists argued that early differences in socialisation experiences and environmental factors might better explain differences in cognitive ability (nurture). More recently, the limitations of both the nature and nurture perspective in isolation have being realised (Halpern & Tan, 2001; Priess & Hyde, 2010). Increasingly, researchers acknowledge the need for a more comprehensive theoretical framework that encompasses biological, social and psychological factors, which Halpern (2000, 2011) has termed psychobiosocial models of sex differences. Such models are a network of multiple levels of input featuring proximal causes such as biology and early socialization experiences as well as more distal causes such as
cultural influences and sex-role stereotypes (Wood & Eagly, 2002). Such a perspective may also help explain individual differences in the development of cognitive ability, and the reasons why some males and females perform at lower or higher level than their same-sex peers. As noted by Hyde (2005), there is greater within-gender variability than between-gender differences, an observation first made by E. L. Thorndike (1914) who argued for the importance of understanding individual differences factors. Investigating individual difference factors around gender (such as sex-role identification, endorsement of gender stereotypes, etc.) may shed light on the mechanisms involved, which can inform the development of educational interventions.

A consensus statement by researchers in the area of cognitive sex differences is that there is a need for both basic and applied research to improve educational outcomes for boys and girls (Halpern et al., 2007b; Neisser et al., 1996). Thus, the second aim of this programme of research was to investigate the contribution of an individual differences factor, sex-role identification, on the development of sex-typed cognitive abilities. Nash (1979) had hypothesised that sex-role identification acted as a mediator
of sex differences in intellectual functioning leading to the improved acquisition of visual-spatial and verbal abilities (see Figure 1.1).

**Figure 1.1.** Nash’s (1979) sex-role mediation hypothesis. Contribution of sex/gender is mediated by sex-role identification, with masculine sex-roles promoting the development of visual-spatial reasoning and feminine sex-roles encouraging the development of verbal abilities and general language proficiency.

A meta-analysis by Signorella and Jamison (1986) of studies investigating the sex-role mediation hypothesis found support for visual-spatial ability, but a lack of empirical studies that actually tested verbal abilities prohibited drawing any conclusion for language. In the passage of time since, society has seen considerable change in the status of men and women (Auster & Ohm, 2000), including cultural prescriptions for
sex-roles. This merits further investigation, to test whether the association between sex-role identification and verbal/visual-spatial abilities is still found in modern samples. Similar predictions are also made by Eagly and Wood’s (1999) social role theory (see Section 2.3.3.1), which posits that psychological sex differences in thought and behaviour arise from the segregation of society into masculine and feminine sex-roles.

1.3. Importance of sex difference research in educational psychology

Equality of educational outcomes is a desirable social good – if one or more subgroups in society lag significantly behind the majority, this may result in deleterious outcomes not just for individual members of that group (reduced job security, lower socioeconomic status) but also for society as a whole (entrenching social disadvantage). Indeed a fundamental tenet of egalitarian societies is equality of education outcomes, and the issue of sex differences in education has been given considerable attention by psychological researchers, educators, parents and policy makers (Dwyer, 1973; Halpern, 1997; Halpern et al., 2007b; Newcombe et al., 2009).

There are two central educational issues that highlight the importance of additional research in this area. The first is the underrepresentation of women in science, technology, engineering and mathematics (STEM)-related fields and STEM literacy. The second is the pattern of lower educational aspirations and achievement of men beyond compulsory schooling.

1.3.1 Underrepresentation of women in STEM fields, and STEM literacy

The skills required for the workforce have substantially changed in recent decades. Whereas once a basic proficiency in reading, writing, and mathematics were seen as important educational milestones, in recent years, advanced mathematical, and scientific literacy have been seen as important skills for occupational success. Additionally, continued scientific progress depends largely on the quality of the pool of
future scientists, and inspiring new entrants to pursue science and technology is critical for maintaining a competitive and growing economy (President’s Council of Advisors on Science and Technology, 2010). Current shortfalls in the number of STEM graduates required for industry are currently mitigated to some degree by importing expertise from abroad (e.g., the H1-B visa in the United States, and the 457 visa for Australia), but the need to broaden the base of the STEM-ready workforce has long been acknowledged by the scientific profession, including the numbers of women participating in scientific and technology research. Though modest progress has been made towards closing gaps in recent decades, women continue to be underrepresented in their participation in STEM-related fields in most Western nations including the United States (National Science Foundation, 2017), Australia (Bell, 2010), as well in large parts of the developing world (Sugimoto, Larivière, Ni, Gingras, & Cronin, 2013). In the Australian context, this concern is matched by government policy action - considerable efforts have been made by the Australian Federal Government in recent years to address gender inequalities in STEM opportunities and labour supply (outlined in Latimer et al., 2019). Despite making some progress in recent years, globally women make up only 28.8% of the scientific research workforce and less than a third in Western nations. The greatest gender imbalance starts at tertiary and postgraduate studies, which has been termed the “leaky pipeline” problem. At every step after compulsory schooling (college/undergraduate, postgraduate, PhD, etc.) women become rarer and rarer, and when they leave study women often do not return (Alper & Gibbons, 1993). There are, however, exceptions to this general rule: women comprise approximately 70% of graduates at each degree level for psychology, and there are slightly more women than men in life sciences (Hanson, Schaub, & Baker, 1996; National Science Foundation,
This distinction is important to note, because it suggests that it is not due to other social factors such as career-family pressures, or lack of ability or interest.

The question of the underrepresentation of women is complex and yet to be fully understood (Ceci & Williams, 2011; Halpern, 2007), but many researchers have argued that the early sex differences in mathematics and science achievement at school may play a role. Although the effect size for mean sex differences in mathematics and science may be small, Hyde et al. (2008) has noted that larger sex differences are found for more complex problem solving tasks during the high school years (Brunner, Krauss, & Kunter, 2008). Researchers have also observed that there is a pronounced disparity in the sex ratio of high achievers in mathematics and science who are more likely to go on to pursue further STEM-related studies (Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Reilly, Neumann, & Andrews, 2015). Additionally, gender stereotyping associating science with masculinity helps to shape girls’ attitudes towards mathematics and science, resulting in reduced mathematics and science self-efficacy beliefs during adolescence and young adulthood (McGraw et al., 2006; Reilly, Neumann, & Andrews, 2017). A large body of research by Eccles (1994, 2007) and colleagues have shown that attitudes towards science and mathematics, cultural sex-role stereotypes, and individual self-efficacy beliefs guide students either towards (or away from) the pursuit of STEM careers (Eccles, 2013; Else-Quest, Mineo, & Higgins, 2013; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002).

While the importance of increasing the representation of women in STEM professions has been argued as an economic and logistical issue for industry, there are further reasons to see it as a desirable social good. Affording women the opportunity to pursue a STEM-related career if they so choose is also important as a matter of wage equity - wages in these traditionally male-dominated professions are typically much
higher than the median wage and offer greater job security (Hill, Corbett, & St. Rose, 2010). For example, in the United States, STEM workers earn 29% more than their non-STEM counterparts, and growth in STEM occupations exceeds growth in non-STEM occupations (Office of the Chief Economist, 2017). Similar findings of employment growth are found in Australia (Office of the Chief Scientist, 2014). Providing equality of educational outcomes in science and mathematics is therefore an issue of gender equity (Halpern et al., 2007b; Hyde & Lindberg, 2007), and can ensure that important subgroups (such as women) do not get left behind. The skills required for the workforce have substantially changed in recent decades. More advanced mathematical and scientific literacy are now seen as important skills for occupational success even if one does not seek to become a scientist or computer programmer.

Science literacy is also important for full participation in society. For example, attaining a basic science literacy is important for understanding health-related information (such as leading healthy lifestyles, or the need for preventative vaccination), as well as important social and political issues that impact on society (such as the science of climate change, need for public funding of space research, or any one of a myriad of other issues where STEM and public policy intersect). Given that women are often more influential than men in family medical decision-making and health information seeking (Abrahamson, Fisher, Turner, Durrance, & Turner, 2008; Beier & Ackerman, 2003; Washington, Burke, Joseph, Guerra, & Pasick, 2009), a basic proficiency in understanding of scientific and medical issues is also important for a healthy society. For example, research suggests that poorer health literacy is associated with a reduced likelihood to undergo routine preventative screening and poorer treatment outcomes (Vahabi, 2007). Sex differences researchers often stress the importance of increasing the representation of women in STEM-fields as an
Sex and sex-role differences in cognitive abilities

occupational/economic issue, but traditionally have placed less emphasis on the social consequences of reduced STEM literacy in women.

1.3.2 Sex differences in literacy, schooling, and entry to higher education

The second issue concerns lower educational outcomes for males (reading and writing literacy, completion of schooling and pursuit of tertiary education). Historically women had lower educational attainment than men for the early half of the 20th century due to societal barriers (Alexander & Eckland, 1974). From the 1960’s onward changes in societal attitudes towards the status of women saw a rise in the number of women completing high school and seeking further education across most developed nations. In more recent decades, the pattern has completely reversed – boys are more likely to drop out of high school before completion than girls (85% versus 78%) (Table A2.4: OECD, 2011), including in Australia where the sex differences in Grade 12 completion rates has now reached 10% (Marks, 2008). Women now significantly outnumber men in attending college education in the United States (Conger & Long, 2010), and similar patterns are found internationally. For example in the context of Australia, since 1985 more women than men have entered tertiary studies each year. The trend appears stable, if not slightly widening (see Figure 1.2). Once enrolled, males have a significantly higher dropout rate in their first year of study and lower overall completion rates (70.9% versus 75.5%), as shown in a cohort analysis of Australian students from 2005 through to 2013 (Department of Education, 2014). Across OECD nations, far more females than males enrol in further tertiary education, with the only three exceptions being Switzerland, Turkey and Japan (OECD, 2016). Jacob (2002) notes that this increases for low-income and minority students where women are 25 percent more likely than men to enrol in tertiary education.
Compounding the issue of disparities in educational attainment, there are also pronounced sex differences in reading literacy (Hedges & Nowell, 1995; Lynn & Mikk, 2009; Mullis, Martin, Kennedy, & Foy, 2007), grammar (Stanley et al., 1992) and writing skills (Reynolds, Scheiber, Hajovsky, Schwartz, & Kaufman, 2015). A full appreciation of the extent of the male-female gap in reading and writing can be gained by considering the effect size for reading and writing. While sex differences in mathematics and science are typically small in magnitude by Cohen’s (1988) effect size guidelines, sex differences in reading and writing achievement typically fall in the medium to large range. But many of the datasets used in these analysis are dated, and further research is needed with modern samples (see Chapter 5). But there is tentative
evidence that this reading gap remains. In a recent educational assessment of reading literacy attainment in OECD nations (PISA 2012), girls outperformed boys in reading on average by the equivalent of a full year of schooling (OECD, 2015; Indicator A10).

Unlike the rise in women’s educational aspirations, the issue of sex differences in reading and writing literacy is not a recent phenomenon – in a systematic review of the research literature available at the time, Maccoby and Jacklin (1974) noted that sex differences in language were ‘firmly established’ (p. 351). Nowell and Hedges (1998) reviewed several decades (1960-1994) of nationally representative testing data for students in the U.S.A., finding the gender gap in language proficiency (reading and writing) had remained stable over a 34 year period. That there was no change (either increasing or decreasing) is an important consideration for educators, because it demonstrates that improving the educational aspirations of girls and women has not come at the cost of boys’ academic achievement.

Just as the lowered educational aspirations of girls and women were once an important target for intervention as a matter of gender equality, a number of educational researchers have expressed concern that the combination of poorer language development in reading and writing skills and a pattern of lower educational aspiration in boys and men merits educational intervention (Alon & Gelbgiser, 2011; Buchmann, DiPrete, & McDaniel, 2008; Entwisle, Alexander, & Olson, 2007). Childhood reading ability is also a strong predictor of eventual adult socioeconomic status, even after controlling for the effect of birth SES (Ritchie & Bates, 2013). Sex differences in educational achievement can be seen from primary school and continuing through to high school, with girls achieving higher grades than boys, including in mathematics and science (Duckworth & Seligman, 2006; Perkins, Kleiner, Roey, & Brown, 2004). Well-developed literacy skills are essential for academic success across all levels of
compulsory schooling (Dockrell, Lindsay, & Palikara, 2011). For those students who do enter tertiary education, there is also a significant female advantage in coursework grades and GPA (Duckworth & Seligman, 2006; Perkins et al., 2004). In a recent meta-analysis of scholastic achievement, Voyer and Voyer (2014) reported a significant sex difference between male and female students in college and university of $d = .21$ 
$[95\%CI = .17$ to $.25]$, which is a small effect size but exceeds Hyde and Grabe’s (2008) critical value for nontrivial sex differences by a factor of two.

The issue of lowered educational expectations and educational attainment for males is a complex and contentious issue, with a wide variety of non-cognitive and social factors contributing to the gender gap. However, females do approach tertiary education with a substantial advantage over their male peers in reading and writing proficiency (Hedges & Nowell, 1995; Lynn & Mikk, 2009) and it has been argued that sex differences in reading and language proficiency may be at least partially responsible for lower commencement and completion rates (Buchmann & DiPrete, 2006).

Successful tertiary education requires a variety of skills, including the ability to read and comprehend written material such as textbooks, readings, scientific papers and other documents. It also requires students to write essays and reports, which form part of student assessment. If male students enter tertiary education and training without fully developing their language competency, it could have a deleterious effect on educational success.

In addition to tertiary education, reading and writing are important skills in their own right. Regardless of whether a student decides to pursue tertiary studies, seek a trade qualification, or enters the workforce directly, educators and policy-makers see value in citizens attaining reading and writing literacy for full participation in society. While there are manual jobs and trade professions that do not require such skills, in the
future increased automation and disruptive technological change may reduce or eliminate the need for unskilled or lowly-skilled labour (Muro, Maxim, & Whiton, 2019). Increasingly economists and public policy makers see automation as a gendered issue, as those professions most likely to be automated (e.g., manufacturing, assembly, driving) are disproportionately male-dominated, while occupations that are more resistant to the threat of automation (e.g., nursing and medicine, child- and aged-care) are largely female-dominated (AlphaBeta, 2017). This means that those males who are without higher reading and writing skills may encounter difficulties reskilling and pursuing tertiary education or seeking retraining if required. Economic predictions of labour market trends predict dramatically higher male unemployment as a result of automation (Bloom, McKenna, & Prettner, 2018; Muro et al., 2019), but improved literacy would improve opportunities for retraining in new skills. I would argue that reducing or substantially eliminating sex differences in reading and writing is an important target as a matter of gender equity and social cohesion.

1.3.3 Summary of educational importance

Systemic disparities in educational achievement can have profound consequences for men and women’s lives beyond their schooling (Priess & Hyde, 2010; Riegle-Crumb, 2005). In the United States, for example, gender equity in educational outcomes is mandated by Title IX of the Education Amendments Act of 1972, and has led to considerable efforts to increase the number of girls studying mathematics and science classes at high school (Walters, 2010), as well as substantial funding of basic and applied research. Other equality of educational outcomes legislation requires governmental agencies like the National Science Foundation to collect annual data on the number of women starting and completing postgraduate training in a STEM field, as well as those entering and leaving the workforce (National Science Foundation, 2017),
so as to track whether educational initiativestranslate into real world outcomes. Similar initiatives can be found internationally (UNESCO, 2012), making the underrepresentation of women in STEM a high profile research issue. However, the issue of lowered educational aspirations for males (as well as poorer reading and writing skills), receives a comparatively less attention by researchers at a time when many male-dominated occupations are threatened by automation. Both issues (women in STEM, men’s educational aspirations and reading/writing literacy) are important targets for further study and worthy in their own right.

Gender gaps cross all strata of society and have may impact the life outcomes of a significant portion of society (Buchmann et al., 2008). In a debate on the merits of conducting and publishing sex difference research in American Psychologist, Eagly (1990, p. 562) has called the scientific debate over gender differences “one of the most important scientific debates of our time”, while Halpern, Beninger and Straight (2011, p. 266) argue that furthering understanding of sex differences is “crucial” to improving educational achievements and aspirations for both genders. Though cognitive ability alone is but one of many factors associated with educational success, it is highly tied to academic self-efficacy beliefs and motivation, and guides adolescent and young adult career decision-making processes especially for stereotypically gendered professions (Eccles, 2013). But sound educational and public policy decisions require sound empirical evidence (Halpern et al., 2011), and much of the literature in this area is dated.

1.4 Research questions

There are important research questions about sex differences that need to be addressed to advance the field theoretically and to provide the necessary information to guide educational and wide society policies. As mentioned there is considerable debate
in the literature over fundamental issues such as the magnitude of sex differences in cognitive abilities. This has proved difficult to answer definitively for the reasons outlined earlier, including small sample sizes and selection of samples that are not representative of the general population. On this subject, Halpern (1989) once noted that what researchers see depends on where (and how) they look. As a result, the present literature is shaped by the choices and ideology of the researchers in the field. For example cognitive sex differences tend to be larger in adolescence and young adulthood, so testing for sex differences only at earlier ages serves to bolster evidence for the null hypothesis (while older-aged samples that might have revealed meaningful differences were sometimes not examined). Additionally, the vast majority of sex difference research focuses on cognitive tasks where females score lower than males (e.g., visual-spatial ability), while the issue of language differences where opposite trends are found (e.g., verbal ability) is less often investigated. Furthermore if sex differences in cognitive ability are even partly the product of psychosocial factors (such as the roles of women and men in society), and if these have changed with the passage of time, it raises the question of temporal stability. Even the most carefully selected demographically representative sample is limited to a single cohort in time. There is a need for further basic research to determine the magnitude of cognitive sex differences to address these gaps in the scientific literature.

Despite the identification of sex differences in cognitive ability quite early in the history of psychometrics, relatively modest progress has been made in explaining why such group differences emerge at the population level. A variety of explanations have been proposed (see Section 2.3) but three areas that seem promising are the contribution of sex-role identification to cognitive performance, the role of situational factors (such as appraisal of test content, or knowledge of gender stereotypes), and sex differences in
self-estimated intelligence (if one believes they are lower in intelligence than their peers, it may become a self-fulfilling prophecy). Accordingly, four research questions (RQ) have been identified for this research project. These are listed below followed by a short outline.

1.5.1 RQ 1: Magnitude of sex differences in cognitive abilities

The first research question aimed to examine the sex differences hypothesis in modern samples, and to provide an estimate of the magnitude of effect sizes for verbal ability and quantitative ability. This hypothesis was tested by examining archival data in language usage (reading and writing) and quantitative reasoning (mathematics and science literacy). By examining archival data collected over a prolonged time period (for example, the National Assessment of Educational Progress, NAEP in U.S. samples) it also allowed for testing of the hypothesis that changes in societal values and the roles of men and women in society have subsequently reduced or eliminated sex differences in cognitive ability. The question of whether there are still observable sex differences in visual-spatial ability in modern samples has been convincingly demonstrated by meta-analytic reviews of spatial ability (Voyer et al., 1995); no further research on this matter is needed at the present time. Examining large international assessments of student achievement (e.g., OECD’s Programme for International Student Assessment, PISA) also allows for investigation of cross-cultural patterns, to determine the portion of sex differences that arise independent of sociocultural factors.

1.5.2 RQ 2: Contribution of sex-role identification to cognitive performance

The second research question concerned the contribution of sex-role identification to the development of sex differences in cognitive ability. Specifically, it examined Nash’s (1979) sex-role mediation theory. This theory predicts that masculine personality traits are associated with greater visual-spatial ability, and that feminine
personality traits are associated with greater verbal and language ability (Dwyer, 1974; Nash, 1974). Androgynous subjects (high masculinity and high femininity) should score highly on both visual-spatial and verbal tasks. A meta-analysis of research studies had shown support for the theory (Signorella & Jamison, 1986), but the authors note that included studies were subject to a number methodological limitations including sample size and breadth of tasks investigated. Halpern (2000) reviewed support for the sex-role mediation hypothesis in some length, nothing that while there was an initial spurt of promising research the hypothesis had “not held up well” (p. 243) in subsequent decades (at least for language tasks). Given the passage of time, it might also be the case that the theory lacks predictive validity for modern samples. Before recruiting subjects for the experimental study (Chapter 8), a meta-analytic review (see Chapter 7) of empirical studies (both published and unpublished) was conducted to examine the association between masculine personality traits and visual-spatial ability in modern samples published since the Signorella and Jamison (1986) review.

1.5.3 RQ 3: Contribution of situational factors to cognitive sex differences

The third research question concerned the role of the testing environment (such as task instructions) and stereotype threat on cognitive performance (investigated in Chapter 9). The approach sought to experimentally manipulate participant’s perceptions of the test content as being either masculine or feminine in nature by changing task instructions and providing additional information. The sex-role identification held by participants was also measured. The methodology thus allowed for testing the separate and joint effects of sex-role identification and stereotype threat.

1.5.4 RQ 4: Contribution of sex-role identification to self-estimated intelligence

The fourth and final research question concerned sex differences in self-estimated intelligence (SEI). As noted earlier, males and females perform equivalently
on psychometrically measured intelligence ($g$) and IQ (Halpern, 2000; Jensen, 1998). However males report significantly higher perceptions of their own intelligence than do females. The fourth research question sought to determine whether sex differences in self-estimated intelligence might also be explained by sex-role identification; specifically that masculinity would be associated with higher self-estimated IQ, while femininity would be associated with lower self-estimated IQ. This is reported in Chapter 10.

1.5 Overview of current research

Chapter 2 outlines a literature review of theoretical perspectives on the development of sex differences in specific cognitive abilities, as well as a summary of empirical research findings for verbal and language abilities. Chapter 3 presents a literature review specifically on visual-spatial reasoning. From thereon, the current programme of research was divided into two sections. The first section outlines a series of meta-analyses examining archival data on student testing data from large nationally-representative samples in the United States (National Assessment of Educational Progress; Chapters 4 and 5), and internationally from Programme for International Student Assessment (PISA; Chapter 6), to determine the magnitude of sex differences in reading, writing, mathematics and science achievement. The second section contains empirical studies, investigating the contribution of sex-role identification to sex-typed cognitive abilities (visual-spatial and verbal abilities), sex-role conformity pressures, and finally to self-estimated intelligence scores. Given the passage of time since Nash’s (1979) sex-role mediation hypothesis was conceived, it is possible that changes in sex-role norms and gender stereotypes might have rendered it outdated, or that past research findings might not replicate to modern cohorts of students. Before recruiting participants for the primary empirical study, it was deemed prudent to conduct a meta-
analysis of research findings on the contribution of masculine sex-role identification to visual-spatial ability (Chapter 7). The lack of sufficient studies investigating the second part of the sex-role mediation hypothesis (feminine sex-role identification and verbal abilities) precluded conducting a similar meta-analytic review.

Chapter 8 presents an investigation into the sex-role mediation hypothesis. While a considerable number of studies have sought to test the hypothesis previously, they have been hampered by substantial methodological limitations. These include i) inadequate sample sizes and low statistical power (Hansen, Jamison, & Signorella, 1982), ii) employing ad-hoc, peer-rated, or psychometrically invalid measures of sex-role identification (an issue addressed further in Signorella & Jamison, 1986), iii) administering tests to one gender only (e.g., Newcombe & Dubas, 1992), iv) examining associations between sex-roles and visual-spatial ability but neglecting to measure verbal abilities, and v) considering only one type of visual-spatial ability (e.g., mental rotation) rather than a broad range of tasks (i.e., spatial perception, spatial visualization). The research reported in Chapter 8 aimed to address these limitations from previous research.

Chapter 9 presents an investigation into the joint effects of sex-role identification and situational factors (such as task instructions, and knowledge of gender stereotypes) on cognitive performance on two visual-spatial tasks. The final empirical study considers another important psychosocial factor in the development of sex differences in cognitive abilities, that of intellectual self-concept. Chapter 10 investigates sex differences in self-estimated intelligence, an important contributor to personal self-efficacy and intellectual functioning. Recall that men and women as a group do not differ in general intelligence, but a considerable number of studies have identified that men self-report their intelligence to be significantly higher than women
(termed the male-hubris female-humility problem by Furnham, Hosoe and Tang, 2001).

Chapter 10 aimed to test whether masculine or feminine sex-role identification might offer an explanation for the apparent gender gap in self-appraisal of intellect.

Finally Chapter 11 contains a general discussion of the implications of the empirical studies presented in this thesis, along with the findings from archival research into patterns of sex differences outlined in the meta-analyses. From this, an updated psychobiosocial model of sex differences is presented. The model incorporates new findings on sex-role identification, intellectual self-concept, as well as broader macro-level cultural contributions such as gender segregation and inequality.

References


Dockrell, J., Lindsay, G., & Palikara, O. (2011). Explaining the academic achievement at school leaving for pupils with a history of language impairment: Previous academic
achievement and literacy skills. Child Language Teaching and Therapy, 27(2), 223-237. doi:10.1177/0265659011398671


SEX AND SEX-ROLE DIFFERENCES IN COGNITIVE ABILITIES


**Chapter 2 – Literature Review**

This section presents an overview of research into sex differences, divided into three parts. Firstly, it offers an overview of empirical evidence of sex differences in specific cognitive abilities reported in the literature and how this evidence has changed with the passage of time. For ease of reference, a summary of this material is presented in Table 2.1, including estimates of effect size strength. Secondly, it presents an overview of popular beliefs by laypeople about the nature of sex differences in intelligence of the sexes which often differs from reality. Finally, it reviews theoretical perspectives on the origins of sex differences in cognitive abilities.

**2.1 Summary of Research Findings**

As outlined earlier, males and females are equivalent in overall intelligence. When representative samples of men and women are compared at a population level (e.g., Deary, Thorpe, Wilson, Starr, & Whalley, 2003), their psychometrically assessed general intelligence (IQ) is equivalent (Colom, Juan-Espinosa, Abad, & García, 2000; Jensen, 1998; Neisser et al., 1996). Sternberg (2014, p. 178) concluded that “there is no evidence, overall, of sex differences in levels of intelligence” while Halpern (2000, p. 218) concluded that “sex differences have not been found in general intelligence”. However a large number of studies have noted that the performance of males shows more statistical variance (i.e., there are a greater number of high- and low-achieving males at the extreme tails of the distribution), termed the greater male variability (GMV) hypothesis (H. Ellis, 1904; Feingold, 1992; Shields, 1982). This means that even if there were no mean sex differences in intelligence, recruiting from highly selective samples that are high in ability (such as college student subject pools, or a group of gifted and talented) may show higher male intelligence scores. Thus the reader is cautioned against a small number of contrarian studies (e.g., Irwing & Lynn, 2005)
showing purported sex differences in general intelligence, but which are methodologically flawed in the representativeness of their sample selection.

Nonetheless, while there negligible sex differences in psychometrically assessed intelligence, patterns of sex differences are frequently observed for more specific cognitive abilities (i.e., some types of cognitive tasks show a female advantage and others show a male advantage). Maccoby and Jacklin (1974) conducted the first rigorous review of the available sex difference literature. Their review concluded that there were robust sex differences for three types of cognitive tasks: verbal abilities, visual-spatial abilities, and quantitative reasoning (mathematical and scientific reasoning). Each are reviewed herein. Subsequent researchers such as Halpern (2000) have also identified sex differences for other types of cognitive abilities such as memory (Herlitz & Rehnman, 2008), moral cognition (Jaffee & Hyde, 2000; You, Maeda, & Bebeau, 2011), nonverbal perception (Hall, 1978; LaFrance & Vial, 2016), and emotional intelligence (Petrides & Furnham, 2000) so the reader is cautioned that a tripartite classification of sex differences is not exhaustive. But as these traits involve other non-cognitive processes (e.g., social, emotional), they are beyond the scope of this review which is confined to intellectual functioning. A review of sex differences in memory is also presented as memory (particularly working memory) affects performance on other cognitive tasks (Daneman, 1991; Peng et al., 2017; Schmader & Johns, 2003)

2.2.1 Verbal abilities.

In a classical text on differential psychology, Anastasi (1958) concluded that females were superior to males in language ability from infancy throughout adulthood. Halpern (2000, p. 93) argued that “evidence from a variety of sources supports the findings that, on average, females have better verbal abilities than males”. While the
extent of sex differences in other types of cognitive ability may be frequently disputed, sex differences in language and verbal ability are generally regarded as “fairly well established” (Maccoby & Jacklin, 1974, p. 351), while Hyde and Linn (1988) acknowledged “there is a clear consensus” (p. 53) to the point where it has become one of the few generally accepted truisms in sex difference research (L. Ellis et al., 2008; Galsworthy, Dionne, Dale, & Plomin, 2000; Kimura, 2000). But just what types of task constitute verbal ability, and do males and females differ on all language tasks?

Halpern and others have argued that verbal ability is not a unitary construct (J.B. Carroll, 1941; Halpern, 2000), and instead encompasses a wide range of tasks rather than a single underlying ability. Unlike other types of cognition such as visual-spatial ability there exists no widely accepted formal criteria for defining and organising verbal and language abilities. Even the term ‘verbal abilities’ is a misnomer because the commonly accepted lay understanding of verbal pertains to spoken utterances rather than written ones, which is why my preferred terminology (adopted herein) is verbal and language abilities. The term applies to any task involving words and language, covering both language reception (listening and reading) and language production (orally or in written form). Thus it encompasses a broad constellation of tasks, including verbal fluency, grammar, spelling, reading and writing, oral comprehension, speech production, vocabulary, as well as related tasks such as synonym generation and verbal analogies. By convention these tasks are generally conducted in a speaker’s native language (which may not always be English, and hence sex differences are not confined to a single language or culture), but there is also a body of evidence demonstrating greater female proficiency for second language acquisition. Additionally significant sex differences have been found for pictographic tasks that require reading novel/unfamiliar characters such as the Digit-Symbol Test. This is commonly regarded
as a measure of coding/processing speed, but also involves activation of language regions and phonological skills (Majeres, 2007; Royer, 1978).

Maccoby and Jacklin (1974, p. 351) concluded that females tend to score higher than males on all verbal ability tasks. They wrote “Girls score higher on tasks involving both receptive and productive language, and on ‘high-level’ verbal tasks (analogies, comprehension of difficult written material, creative writing) as well as upon the "lower-level' measures (fluency). The magnitude of the female advantage varies, being most commonly about one-quarter of a standard deviation.”

As was the practice at the time, Maccoby and Jacklin produced what has been termed a narrative review as the statistical techniques of meta-analysis were not yet developed. Hyde and Linn (1988; Table 3) conducted a subsequent meta-analysis of research into verbal ability, finding a slight but non-trival sex difference favouring females of $d = .11$ across all ages, and $d = .20$ in adults. However there was considerable variation in magnitude and direction across types of tasks. Females performed better on speech production ($d = .33$), anagrams ($d = .22$) and general language ability across mixed tasks ($d = .20$), but did not significantly differ in vocabulary and there were only negligible differences in reading ($d = .03$) and writing ($d = .09$). Males, however, did do significantly better than females on tasks of verbal analogies ($d = .16$) which differs from Maccoby and Jacklin’s conclusion. Hyde and Linn concluded, somewhat contentiously, that such evidence demonstrates that “differences in verbal ability no longer exist”, p. 53.

The technique of meta-analysis can be used to aggregate research findings and objectively measure the size of experimental effects, giving greater credibility to the conclusions drawn than traditional vote-counting or narrative literature reviews (Hedges & Olkin, 1985; Rosenthal, 1984), such as those reported by Maccoby and Jacklin
(1974). It is limited, however, by the selection of studies that comprise the dataset and whether it represents a complete view of the research literature (Rosenthal & DiMatteo, 2001). Selective inclusion or omission of important studies (“cherry-picking”) will therefore lead to a biased interpretation but which carries the air of legitimacy. Given that verbal abilities is a broad constellation of tasks with a nebulous definition, it is entirely reasonable that some tasks be excluded. However, in a critique Stumpf (1995) argued that Hyde and Linn’s meta-analysis is characterised by several important omissions: namely the tasks that Maccoby and Jacklin had identified as showing larger sex differences, such as writing, grammar/language usage and verbal fluency. By way of example, Hyde, Geiringer and Yen report an extremely large effect size ($d = -.76$) for a verbal fluency task, but Hyde and Linn chose not to include that verbal fluency task in their analysis.

Additionally a number of large nationally representative samples of reading and writing reported by Hedges and Nowell (1995) which fell in the time frame defined by Hyde and Linn for analysis (up to 1986) were also excluded without offering any justification. Their inclusion would have dramatically increased the overall effect size calculation, as well as the effect size for reading and writing abilities. Therefore Hyde and Lynn’s meta-analysis is likely to be a lower-bound estimate of the true effect size in verbal and language abilities, and runs contrary to the conclusions drawn by other prominent researchers in this field (David C Geary, 1998; Halpern, 2011; Hedges & Nowell, 1995; Kimura, 2000; Miller & Halpern, 2013; Neisser et al., 1996). However if there are meaningful differences between males and females in language ability then this should be reflected in other objective measures of performance such as student grades in English and other language classes. A subsequent meta-analysis conducted by Voyer and Voyer (2014) examined this question by investigating sex differences in
student grades. Based on an analysis of 81 studies, the authors report that females earned significantly higher grades in English and language classes, with a small to medium effect size ($d = .37$).

### 2.2.1.1 Developmental differences.

When do sex differences in language and verbal abilities appear, and do they persist into adulthood? All healthy children undergo the same developmental trajectory, from early cooing and babbling as an infant, to the beginnings of speech and repetition of heard words as a toddler around 12 months of age, followed by an explosive growth in vocabulary from 12-24 months after which vocabulary growth becomes linear. Sex differences in language have been observed from infancy, with newborn girls beginning cooing and babbling sooner, and engaging in greater vocalization (Balint, 1948; Gatewood & Weiss, 1930; Korner, 1973). Fenson et al. (1994) report that between 8 and 16 months infant girls understood more words than boys while Halpern (2000) notes that onset of speech is on average about 1 month earlier in girls than boys. Girls also show an early advantage in vocabulary size. In a study on vocabulary growth rates in infants, Huttenlocker et al. (1991) found that while there is only a modest 13-word difference between girls and boys at age 16 months, this quickly expands to a 51-word difference by 20 months and 115-word difference by 24 months. A similar study by Fenson et. al., (1994) also found that after the age of 14 months, infant girls produced more words with greater sentence complexity. This trend was observed until at least 30 months of age when the study ended. When children enter kindergarten, girls already show a significant advantage in reading, and this trend continues throughout primary school (Robinson & Lubienski, 2011) and high school (Lynn & Mikk, 2009). Girls also tend to be better at spelling than boys from early childhood onwards, as well demonstrating greater proficiency in punctuation, grammar and language usage from
primary school at least until high school (Lynn, 1992; D. J. Martin & Hoover, 1987; Reynolds, Scheiber, Hajovsky, Schwartz, & Kaufman, 2015; Stanley, Benbow, Brody, Dauber, & Lupkowski, 1992). There is also evidence that in cognitively healthy older adults, females perform better on tasks involving word definition and verbal learning (Rabbitt, Donlan, Watson, McInnes, & Bent, 1995), verbal fluency (De Dreu, Greer, Van Kleef, Shalvi, & Handgraaf; Lezack, 1995; Maylor et al., 2007; K.W. Schaie, 1996), and general verbal ability (Maitland, Intrieri, Schaie, & Willis, 2000).

Although evidence of sex differences between boys and girls in child samples may seem compelling, the most frequently invoked refutation is that they reflect different rates of maturation (i.e., precocious language development for girls and a developmental delay for boys but ultimately they will reach the same standard). For example, in early childhood girls have a somewhat larger vocabulary than boys, but Maccoby and Jacklin (1974) claimed that parity is achieved by age 3. However, for many language tasks there is evidence of continuity into later development so simple differences in rates of maturation may be discounted if sex differences persist into adulthood. For example, a nationally representative study of adult literacy in the United States found that the sex difference observed in reading and writing in children endures into adulthood with males demonstrating poorer reading and writing proficiency (Kutner et al., 2007). Similar findings are reported below for other language tasks such as verbal fluency, punctuation and grammar. Indeed Hyde and Linn (1988) found a developmental trend across all language tasks surveyed (subject to the limitations about underestimation noted earlier), with quite small sex differences observed during early childhood but an increase into adulthood ($d = .20$ across all tasks). So there is little support for the differential rates of maturation argument as an explanation for apparent sex differences in language.
2.2.1.2 Reading and writing.

One of the most frequently observed differences in language between males and females is the pronounced sex difference in reading ability, but the magnitude can vary considerably across samples depending on demographic factors such as socioeconomic status, ethnicity, rural versus urban sampling (Fenson et al., 2000; Fernald, Marchman, & Weisleder, 2013; Kaufman, McLean, & Reynolds, 1988). Less seldom investigated are differences in writing which tend to be appreciably larger. To quantify this issue, Hedges and Nowell (1995) published a landmark review on sex differences for a range of tasks using a number of large, nationally representative samples, including the National Assessment of Educational Progress (NAEP 1971-1992). They found that girls significantly outperformed boys in the domain of reading in each year of assessment (ranging from $d = -0.18$ to $-0.30$). Furthermore they found that the performance of males was more variable than that of females with an average variance ratio (VR) of 1.12. This resulted in an overrepresentation of boys as poor readers, ranging from 1.5 to 2 males for every female falling in the bottom 10th percentile. Similarly, there were also substantially sized sex differences in writing (ranging from $d = -0.49$ to $-0.55$), as well as greater male variability. This resulted in sex ratios for students falling in the bottom 10th percentile of between 2.6 and 3.3 males to every female (Nowell & Hedges, 1998, p. 38). The magnitude of these effect sizes observed would be inconsistent with Hyde and Linn’s conclusion that sex differences in verbal abilities no longer exist, but as noted earlier were not included in their review.

Sex differences in reading achievement are not confined to particular cultures or nationalities. Unlike sex differences in quantitative reasoning that show substantial variation in size and direction, sex differences in reading achievement are found universally across all nations which would be consistent with a biological contribution. Large international assessments such as the Progress in International Reading Literacy...
(PIRLS) and PISA show appreciable effect sizes favouring females in all countries (Guiso, Monte, Sapienza, & Zingales, 2008; Lynn & Mikk, 2009).

While the evidence for sex differences in reading achievement is strong, relatively few studies have examined the extent sex differences in writing skills, especially in modern samples. Nowell and Hedges (1998) reported a detailed meta-analysis of NAEP writing data from the period 1984-1994, finding substantially sized sex differences in writing favouring females (ranging from $d = -.49$ to $-.55$), and that gender ratios for students falling in the bottom 10th percentile were between 2.6 and 3.3 males to every female (Nowell & Hedges, 1998, p. 38). Although numerous waves of NAEP assessment data have been collected and annual reports note significantly greater female writing performance, there has been no attempt to quantify the extent of sex differences in writing through meta-analysis (an issue addressed in Chapter 5).

Three studies have examined the magnitude of sex differences in writing proficiency by examining standardization data from two educational assessment tools. Camarata and Woodcock (2006) presented data from the normative samples of the Woodcock-Johnson cognitive and achievement batteries, a large representative sample of males and females aged 5 through to 79. Females scored significantly higher in writing achievement, with an average effect size across the lifespan of $d = -.33$. A similar finding was reported by Scheiber, et al. (2015) who analysed a large sample of adolescents and young adults completing the Kaufman Test of Educational Achievement- Second Edition (KTEA-II), which measures participants’ reading, writing, and mathematics. While no difference was found in mathematics, females scored higher than males on the tests of reading and writing ability. The effect size for reading was small ($d = -.18$), but the effect size for writing ($d = -.40$) was twice as large as that for reading. Additionally, Pargulski and Reynolds (2017) present data from the Weschler
Individual Achievement Test- Third Edition (WIAT-III) Written Expression scale, finding a small but non-trivial effect size of $d = -0.25$ with girls scoring significantly higher than boys across all age groups.

While educators are well aware that boys are overrepresented in formal diagnoses of dyslexia and reading impairment (Hawke, Olson, Willcut, Wadsworth, & DeFries, 2009; Limbrick, Wheldall, & Madelaine, 2008; S. E. Shaywitz, Shaywitz, Fletcher, & Escobar, 1990), it is less common knowledge that boys are also overrepresented with writing disorders. For example, in a population-based assessment of all children born in the town of Rochester, Minnesota between 1976 and 1982 that was not subject to a gendered referral bias, Katsuic et al. (2006) found that twice as many boys than girls met the criteria for a clinically significant writing disorder. Berninger et al., (2008) report a strong comorbidity with reading and writing disorders, which suggests that they may share a common neurological impairment.

2.2.1.3 Phonological coding and perceptual speed.

Phonological coding is the ability to produce and manipulate information about the sound structure of verbal stimuli, and to convert between letters and the sounds they represent. Deficits in phonological coding are linked to difficulties in reading words (Vellutino, Scanlon, & Tanzman, 1994). There is evidence that there are sex differences in phonological awareness (Coltheart, Hull, & Slater, 1975). For example, women are faster than men at counting the number of letters in the English alphabet that contain the phoneme ‘ee’, while McGuiness and Courtney (1983) found that women make fewer errors than men when asked to determine whether a target letter is present in word lists delivered orally. Majeres (2007) has argued that an advantage in phonological coding contributes to the sex difference in perceptual speed found on some cognitive tasks, as they are more easily able to convert from symbols to sounds. Consider for example the
digit-symbol substitution task employed in a number of intelligence tests (Jensen, 1998; Wechsler, 1958), which requires subjects to rapidly convert between unfamiliar symbols and numbers. Girls and women are able to read these pictographs and translate them more quickly than men with appreciable effect sizes. Hedges and Nowell found small to medium sized effects (ranging from $d = -0.21$ to $-0.43$), while Lynn and Mulhern (1991) found a medium to large effect on the WISC-R digit-symbol coding task in children $d = -0.69$ and small to medium effects sizes have been found by Feingold (1992) with adults with the WAIS/WAIS-R, $d = -0.34$. A review by Roivainen (2011) reached similar conclusions for the WAIS-III and Woodcock-Johnson instruments.

Another perceptual speed task involving speeded reading is the finding A’s task (Halpern & Tan, 2001; Kimura & Hampson, 1994). Participants are presented with lists of words arranged in columns spread out over several pages, and asked to cross out the letter ‘A’ whenever it is encountered within a time limit. Halpern and Tan (2001) reported a medium to large effect size on this task $d = -0.77$. However many reviews of verbal and language abilities overlook these pronounced differences in speeded reading for novel stimuli and phonological coding.

2.2.1.4 Vocabulary.

Sex differences in vocabulary develop extremely early. Between the ages of 1 and 2 years, girls score significantly higher than boys on a range of vocabulary measures including vocabulary production, sentence length and sentence complexity (Feldman et al., 2000; Huttenlocher et al., 1991). Girls also demonstrate better comprehension of spoken words, and as toddlers their vocabulary comprehension and production expands at a much faster rate than boys until at least 26 months of age (Huttenlocher et al., 1991; Reznick & Goldfield, 1992). Maccoby and Jacklin (1974) claimed though that girls and boys appear to reach parity after two to three years, and so
the research question of sex differences in vocabulary for adolescents and young adults has been seldom investigated. However a careful review of the literature does show some instances where sex differences in vocabulary have been tested and reported. For example, in a cross-sectional study of elementary school students, Gates (1961) reported sex differences in performance on vocabulary tests of girls and boys (grades 2 through 8), which subsequent meta-analysis shows to be a small but non-trivial effect size (weighted $d = -.23, Z = -13.25, p < .001$). Similarly in a large sample of primary and secondary school students Lynn and Wilson (1993) found girls significantly outperformed boys on the Mill Hill Vocabulary Test (MHVT) which measures participants’ ability to provide the meaning of target words. Weschler (1958) reported a small sex difference in performance for the Vocabulary subscale with the original standardization sample of the WAIS (ages 16-64), with females scoring slightly higher than males, $d = -.11$, though cautioned that items were carefully chosen to minimise the extent of sex differences and that it may be an underestimate of the true effect size. Furthermore the meta-analysis by Hyde and Linn (1998, pg. 61) found that in young adults aged 19 to 25, there was a small gender difference in vocabulary favouring girls, $d = -.23$.

Other studies have failed to find meaningful sex differences in vocabulary tests (Storck & Looft, 1973). For example Kaufman, McLean and Reynolds (1988) report no significant sex differences in the vocabulary subtest of the WAIS-R for the American standardization sample of adults, which differed from Weschler’s observations of performance with the earlier WAIS instrument. A similar null finding was reported by Storck and Looft (1973) who reported a detailed analysis of vocabulary word definitions in a cross-sectional sample (ages 6-66). However, in younger samples Lynn and Mulhern (1991) report that boys scored slightly higher than girls on the WISC-R
vocabulary subtest for both the Scottish and American standardization samples, while Kramer et al. (1988; 1997) found a slight but statistically significant difference favouring males aged 5-16. The meta-analysis by Hyde and Linn (1988) across ages found no sex difference in adolescence but curiously significantly greater vocabulary for women in the age category of 19-25 ($d = -.23$). Therefore it is difficult to determine their extent (if any), and further research is required. Presumably these discrepant findings reflect demographic and sampling differences, as well as differences across instruments. Later revisions of the WISC and WAIS were purposively chosen to minimise sex differences so that separate norm tables would not be needed for males and females (Halpern, 2000; Wechsler, 1958).

Vocabulary tests are recognized as a good representative of verbal-intelligence scales, but do not predict achievement in most language tasks (Guilford, 1967). While in younger children vocabulary may be a good measure of language proficiency, in older children and adults vocabulary is less a measure of language ability and more a measure of crystallized intelligence/level of education (where one would not necessarily expect to find sex differences in modern samples). Indeed many studies often use vocabulary as a proxy for crystallised intelligence, as it requires less training and time to administer. At present time there is insufficient evidence to conclude there are sex differences present in vocabulary in late adolescence and adulthood because few studies have examined this question. However larger sex differences are often still found for some tasks that involve access to vocabulary, such as verbal fluency (reviewed below). While absolute vocabulary size and recognition of words may not differ between males and females, it is possible that females are simply more adept at organizing and retrieving lexical knowledge (see below).
2.2.1.4 Verbal fluency

One of the largest reported sex differences in this area is that of verbal fluency (Halpern & Tan, 2001; Hines, 1990), with effect sizes ranging from small to large across samples and type of verbal fluency task. Methodology for verbal fluency tasks varies, but it generally involves asking participants to generate a list of words matching a particular criteria (e.g., words beginning with the letter F, or words that are a type of animal) under strict timing conditions (typically 60 seconds). Although this can be administered orally if individually tested, typically it is administered in a written format suitable for group administration. To date though, there have been no studies investigating sex differences in the format of administration so it is not clear whether the act of writing confers any advantage (such as mental cueing of related words, or differences in writing speed between males and females).

Sincoff and Sternberg (1987) have argued that language performance can be divided into two broad domains – verbal comprehension, and verbal fluency. Though both are equally important for successful verbal functioning, Sincoff and Sternberg note there has been a clear preference by researchers and educators in focusing heavily on assessment of verbal comprehension but seldom investigating, assessing or teaching verbal fluency. Estes (1974) has argued that verbal fluency also involves additional cognitive processes, including strategies to organise their thinking and seek an optimum search strategy. Weiss et al., (2006) have provided functional imaging evidence that men and women may use different strategies for verbal fluency generation tasks and that females use a balance of clustering and switching strategies to maximise word production.

There are three categories of verbal fluency tasks: phonological, semantic, and synonym generation. Phonological verbal fluency measures the ability to produce lists of words beginning with a target letter, and is a commonly performed
neuropsychological task (Lezack, 1995; Tombaugh, Kozak, & Rees, 1999), sometimes referred to as the Controlled Oral Word Association task (COWA; Benton & Hamsher, 1989). Additional restrictions may be imposed to increase the difficulty of the task, such as the requirement that only unique words be produced (fast may be given, but not faster or fastest), and that people’s names and brand/product names are not permitted. This is repeated for several letters (most commonly ‘F A S’) to produce a continuous score, though variations on the stimuli letters are made for non-English languages (e.g., Kosmidis, Vlahou, Panagiotaki, & Kiosseoglou, 2004). Sex differences are frequently found for this type of task. For example, Weiss et al. (2003) found a medium-sized effect ($d = -.45$) in college-aged subjects, while Herlitz et al. (1999) found a similarly sized effect ($d = -.49$). There also appears to be an interaction between sex and years of education. Ruff et al. (1996, Table 3) found a medium effect size ($d = -.53$) in adult samples (16 to 70 years old) who had completed a college education, but a relatively small sex differences ($d = -.10$) across all education levels. Loonstra, Tarlow and Sellers (2001) reached a similar conclusion ($d = -.15$) in a meta-analysis of COWA studies in neurologically healthy adults, noting an interaction with education. This may be reflective of historical effects of female educational attainment of the past, while in modern samples of college students with greater female representation considerably larger sex differences in phonological verbal fluency are typically found. Bucking against this trend though, a large cross-sectional study of adults across the lifespan conducted by Schaie and Hertzog (1983) found substantial differences amongst older adults, as did a later study of neurologically healthy adults ages 35-80 by Herlitz, Nilsson, and Bäckman (1997), $d = -.16$. However, some studies have found only slight or non-significant sex differences (Tombaugh et al., 1999), particularly in extremely young children (Hurks et al., 2006; Regard, Strauss, & Knapp, 1982) as there appears to
be a developmental trend towards larger gender gaps with age and education. Nonetheless in studies with adults, sex differences in phonological verbal fluency are so consistently found that neurologists have proposed separate norms tables for male and female subjects (Benton & Hamsher, 1989; Bolla, Lindgren, Bonaccorsy, & Bleecker, 1990; Loonstra et al., 2001; Ruff et al., 1996).

**Semantic category** verbal fluency tasks require participants to produce words that belong to a particular semantic category (e.g., fruits & vegetables, animals, etc.), and studies typically find much larger sex differences than for phonological verbal fluency. While fruits and animals are the most frequently chosen stimuli as they are accessible to both children and adults (Gladsjo et al., 1999), more difficult versions of the test employ less commonly accessed semantic categories such as types of clothing, professions, games, tools, vehicles, etc. Effect sizes for semantic verbal fluency studies vary depending on the type and age of sample (Klenberg, Korkman, & Lahti-Nuuttila, 2001). For example Gordon and Lee (1986) reported a medium effect size \( (d = -.44) \) for semantic category task in a college-aged sample, while Aceveo et al. (2000) reported similar effect sizes in English \( (d = -.54) \) and Spanish \( (d = -.37) \) samples of healthy older adults. Like phonological verbal fluency tasks there also appears to be a strong interaction between sex and years of education. Additionally Capitani et al. (1999) have claimed though that the choice of stimuli may confer a gendered advantage due to differences in familiarity with those domains; specifically they found that females produced more words for fruits but males produced more words for tools. Thus great care should be taken to select gender-neutral semantic categories. Failure to do so would constitute a methodological flaw in sex difference studies.

**Synonym generation** verbal fluency tasks require subjects to recall the lexical meaning of a word and generate appropriate synonyms, in a similar fashion to the
thesaurus. Such tasks are considerably more difficult than other verbal fluency, as the subject must recall the definition of a target word and spontaneously generate words with similar meanings while inhibiting those that are unrelated. This is also used as a neuropsychological screening tool, referred to as the Controlled Associates Test (Ekstrom, French, & Harman, 1976). Sex differences in synonym generation are extremely large. For example, Hines (1990) reported an effect size of $d = 1.2$ on a synonym generation task with college-aged samples, while Rahman and Wilson (2003) found an effect size of $d = 1.05$ in a sample of men and women. Both these studies showed a substantial female advantage. However, few studies have employed synonym generation tasks and it is not clear whether these represent outliers.

In summary, the magnitude of sex differences in fluency varies considerably depending on the type of sample and whether phonological, semantic, or synonym verbal fluency tasks are employed. In college or university educated samples, effect sizes for phonological and semantic are medium while synonym tasks are large in size.

### 2.2.1.5 Listening skills and comprehension.

One area of verbal abilities that there are few sex differences is listening skills and comprehension. This is an important observation, as serious deficits in listening ability between the sexes would be one pathway by which more general language deficits might arise. Badian (1999) examined the entire population of a school district followed from pre-kindergarten through to grade 7, finding no difference between boys and girls in the SAT listening comprehension subtest in each annual test. Earlier research on listening comprehension by Haberland (1959) similarly found no difference between men and women in college student samples across three types of listening comprehension tests. Keith, Reynolds, Patel and Ridley (2008) examined a cross-section of participants aged 6 to 59 using the Woodcock-Johnson battery of
cognitive tests, finding only minimal sex difference in oral comprehension scores ($d = -0.06$). Though females scored significantly higher in oral comprehension, it falls short of Hyde’s criterion for non-trivial sex differences. Nonetheless, a number of high profile reviews (e.g., David C. Geary, 2010; Kimura, 2000) make the general claim that there are sex differences in listening comprehension. At present time there is insufficient evidence in support of that claim.

2.2.1.6. Verbal learning.

A large body of research has found that there are sex differences in what are termed verbal learning tasks which require subjects to commit to memory word lists (Herlitz et al., 1997; Kramer et al., 1988), with females generally scoring higher on recall in adult and child samples. However females also score higher than males on other types of memory tests such as object location and digit-span tasks (see Section 2.2.4). Therefore it is unclear whether greater female proficiency is the result of employing verbal stimuli, or instead actual differences in working memory. For ease of classification, discussion of sex differences in verbal learning have been represented as primarily a memory task, but some narrative reviews classify it as a distinct verbal ability.

2.2.1.7 Spelling, punctuation and grammar.

The question of sex differences in spelling is most frequently investigated in studies with children where it is a common finding that girls score significantly higher than boys at all grades of schooling (Allred, 1990; Berninger et al., 2008). For example, in samples of students assessed in Australia as part of NAPLAN testing, girls score substantially higher than boys in spelling across all grades (Appendix A3). In the context of educational testing, the most widely administered test of spelling proficiency is the Spelling subtest of the Differential Aptitude Test (DAT; Bennett, Seashore, &
SEX AND SEX-ROLE DIFFERENCES IN COGNITIVE ABILITIES 52

Wesman, 1982), which is a standardised test of educational achievement. Stanley et al. (1992) reported the effect size magnitude from the 1980 standardization sample of the DAT, which was a nationally representative sample of 61,000 American students in Grades 8-12. Girls scored considerably higher than boys in spelling in every high school grade (effect sizes ranging from $d = -.38$ in Grade 8 to $d = -.50$ by Grade 12). An earlier analysis by Feingold (1988) also reports medium-sized differences in spelling on the DAT for earlier standardizations in 1947, 1962 and 1972 with a developmental trend towards wider gender gaps with age. Similarly, Lynn (1992) found significantly higher female performance with the British standardization sample of approximately 10,000 students, with a developmental trend towards larger sex differences with age. Berninger et al. (2008) have noted that the cause of pronounced spelling deficits is not clearly understood or well researched, but that they are frequently observed in students (and their parents) who are diagnosed with dyslexia. Furthermore, childhood spelling problems usually persist into adulthood even if they no longer qualify for a diagnosis of dyslexia (Lefly & Pennington, 1991). Spelling impairments are also related to other writing deficits in composition, as well as handwriting (Berninger et al., 2008).

Another area where males and females have been shown to differ is in their grammatical usage. The above mentioned DAT battery measures punctuation and grammatical knowledge in the Language Usage subtest, which presents subjects with a series of sentences and asks the reader to identify which segment (if any) contains an error. Stanley et al. (1992) report small to medium effect sizes favouring girls (approximately $d = -.40$), which mirrors that found in earlier decades by Feingold (1988) with the same instrument. Greater proficiency in the understanding of correct grammar may also contribute to the observed sex difference in writing noted earlier.
2.2.1.8 Speech production.

As noted earlier, Hyde and Linn concluded that the largest sex difference in verbal abilities was for tasks involving speech production ($d = -0.33$) which is a small to medium-sized effect. As noted earlier, sex differences in speech production are frequently observed in infants with girls speaking earlier and producing more complex speech. Boys are also more likely than girls to exhibit developmental speech delays and specific language impairment (SLI; J. Stevenson & Richman, 2008; Tomblin et al., 1997). The trend towards greater proficiency and usage in speech continues into early childhood. For example, Brownwell and Smith (1973) observed length and complexity of verbalization of speech in four-year-old children across dyads, triads, and small groups, finding that young girls produced substantially more words than young boys (average $d = -0.58$), while similar findings were reported by Busswell (1980) in second and third grade elementary students. A meta-analytic review by Leaper and Smith (2004) of children's talkativeness found that sex differences are present but much smaller in older children however. Furthermore there is little evidence that adult men and women differ in the number of words spoken during a day, despite popular cultural stereotypes and media reports (Elliot, 2009). For example, in a highly cited study Mehl et al. (2007) reported an analysis of daily speech of men and women over a period of 4-7 days, finding no evidence that women were more talkative than men. A subsequent meta-analytic review of similar studies by Leaper and Ayers (2007) found that contrary to prevailing gender stereotypes, men actually spoke slightly more words in a day than women. However the authors concluded that the difference was negligible in magnitude, so presumably the sex differences in amount of speech production observed in children declines during adolescence or adulthood, presumably due to social factors.

There are also sex differences in the quality of speech production reported in the literature. Males are greatly overrepresented in the prevalence of stuttering (Halpern,
1997; Yairi & Ambrose, 1992), though it is not clear to what extent this is the result of referral bias. Halpern (2000) notes that there are three to four times more males than females who stutter. Silverman and Zimmer (1982) also note that when present, the severity of impairment for stuttering is greater in males and may cause more distress. One reason that this may be the case is that females are generally more proficient in speech articulation, which may buffer against speech impairment. In a series of studies on the quality of speech articulation, Hampson and Kimura (1988) and Hampson (1990; 1990) have shown that women articulate more clearly and accurately than men, especially for speeded articulation tasks (e.g., speeded counting, colour reading, and syllable repetition). However, they also found that intra-individual performance is influenced by estrogen hormone levels.

Such findings have been independently replicated by other researchers who have found that women articulate vowels and consonants more clearly and with greater accuracy (Kempe, Puts, & Çárdenas, 2013; Wadnerkar, Cowell, & Whiteside, 2006). Women are also able to read aloud lists of words significantly faster than men and with fewer errors (Majeres, 1999). Sex differences in speech disfluency are also pronounced. Hall (1984) reports that girls and women generate fewer pauses, hesitations, and filler utterances (e.g., umm, err, etc.) in unscripted speech, and are less likely to commit articulatory retrieval errors (e.g., misspeaking the wrong word). These differences in speech were not insubstantial - Hall reported that 9 out of 10 men have more pauses and filler utterances than the average woman, and 3 out of 4 women commit fewer speech errors than the typical man.

2.2.1.9 Second language acquisition.

While most studies focus on sex differences in primary language acquisition, a number of studies have found that females may be better equipped to learn a second
language. In childhood studies girls consistently outperformed boys in studies from the UK (Burstall, 1975), Ireland (Lynn & Wilson, 1993), Israel (Lewey & Chen, 1974, as cited by Lynn & Piffer, 2011), and Sweden (Ekstrand, 1980). Importantly these studies involve objective tests of actual ability rather than awarded grades. Additionally, Payne and Lynn (2011) found a similar female advantage on reading comprehension in college-aged samples from the United States (\(d = -.49\)). On the basis of these and earlier studies, Payne and Lynn concluded that females are better equipped to learn a second language. Females also surpass their male peers in grades for language-related classes during early childhood (Pomerantz, Altermatt, & Saxon, 2002) and adolescence (Mau & Lynn, 2000) and adulthood (Voyer & Voyer, 2014), with a small to medium sized effect \(d = -.37\). However one factor that might influence language acquisition is that girls generally had more favourable attitudes to learning a new language (S. C. Baker & MacIntyre, 2000; Burstall, 1975) and just as with reading, motivation to master a second language may differ.

**2.2.1.10 Verbal reasoning and analogies.**

One exception to the general rule of greater female proficiency for verbal tasks may be for verbal reasoning and analogies. The meta-analysis by Hyde and Linn (1988) found that males actually outperformed females on tasks involving verbal analogies \(d = +.16\), a small but non-trivial effect size. A difficulty in interpreting this result though is that performance on verbal reasoning tasks also requires recruitment of other non-language cognitive processes, such as the routine application of logic and executive functioning. When differences between males and females are observed for such a task, this impurity makes it difficult to determine whether it is due to group differences in verbal and language abilities, or instead in some other cognitive component. For example, Colom et al. (2004) examined sex differences in verbal reasoning in a large
sample of university students by having them complete linear syllogisms exercises (*e.g.*, “Jane is better than Peter, Peter is better than Paul, Who is worse?”). Colom found that subjects develop a mental spatial diagram for solving such analogies. Though a small male advantage on verbal reasoning was found ($d = +.16$), the authors found this difference vanished when statistically controlling for visual-spatial ability (on which males typically perform higher). Thus some tasks of verbal reasoning that appear, at least on the surface, to be dealing with verbally presented information may in fact be drawing on other cognitive processes such as spatial and working memory. Thus any observed sex differences in verbal reasoning should be interpreted cautiously as it may reflect sex differences in other cognitive processes.

Another prominent finding reported in the literature is that males perform slightly higher on the verbal reasoning component of the SAT exam used to assess students suitability for college entry (Feingold, 1988; Halpern, 2000). This is a multiple choice test assessing verbal reasoning and verbal comprehension, but Halpern (2000) notes that it is heavily weighted with analogies which favour males. This finding is consistent across each year of assessment, though a meta-analysis of the effect shows it to be almost trivial in magnitude $d + .0454$ [95% CI = .04 to .05; Appendix A2]. Spelke (2005) has cautioned that conclusions drawn from such a sample provides only tentative evidence because it is not a representative sample: significantly greater numbers of females choose to sit the SAT exams while the male sample is more selective. In order to rule out this possibility, Mau and Lynn (2001) examined data from the Baccalaureate and Beyond (B&B) 1992-1994 longitudinal study, which is a demographically representative sample of American college graduates. They found that males had significantly outperformed females on their SAT Verbal completed before starting college, though the effect size was relatively small ($d = +.14$). Lynn (1994) also
reported a slight male advantage on verbal reasoning tests in standardization samples for the WISC and WAIS intelligence tests in the United States, Europe and China. Taken together with other verbal reasoning studies the evidence shows a slight male advantage on such tasks but it is questionable to what extent such tasks tap actual verbal ability as they are not cognitively “pure”. Additionally there has been one study by Strand, Deary and Smith (2006) that found significantly higher female performance for verbal reasoning in a nationally representative sample of British children ($n = 320,000$, comprising almost half the child population) on the *Cognitive Abilities Test (CogAT) Version 3*, though the effect size was relatively small, $d = -.15$. Therefore it is highly likely that test content and construct validity are a factor, and that sex differences for analogies (if any) are relatively small.

### 2.2.2 Visual-spatial ability.

Sex differences in visual-spatial ability appear to be the most reliably found and have larger effect sizes that most other cognitive tasks (Halpern, 2011), with males scoring considerably higher on average than females for most visual-spatial tasks. Voyer et al. (1995) conducted the most comprehensive meta-analysis to date, finding robust effect sizes and no support for declining sex differences over time. A full review of the literature on visual-spatial reasoning is deferred until Chapter 3, presented as a publication output.

### 2.2.3 Quantitative ability.

Halpern (2000) noted that the term quantitative reasoning is broadly defined as a heterogeneous set of cognitive aptitudes encompassing tasks involving both mathematical and scientific reasoning. Quantitative reasoning also draws on other cognitive abilities, including general reasoning, working memory, visuospatial abilities, semantic memory for facts and concepts, and advanced problem-solving skills (Halpern
et al., 2007). Some authors have used the term quantitative abilities quite narrowly to refer to proficiency with numbers (e.g., calculation), but for the purposes of this review, a more expansive definition encompassing performance in either mathematical or scientific domains is adopted. This differs somewhat from the terminology “Quantitative Ability” ($G_Q$) under the Cattell-Horn-Carrol (CHC) model of cognitive abilities (J. Horn, 1994; J. L. Horn, 1991; McGrew, 1997), as it includes broader scientific problem solving tasks than solely mathematics.

Historically, scientific problem solving was not included in the initial intelligence batteries that CHC theory used for factor analysis, but has been included in other intelligence batteries such as the Armed Services Vocational Battery (ASVAB). Carrol (1993), however, has argued that quantitative reasoning is a broader construct than just computation skill, as it also includes logical reasoning processes “in order to arrive at correct conclusions. The reasoning processes may be either inductive, or deductive, or both” (p. 246). Thus it incorporates elements of logical reasoning and the scientific method. Additionally, Vernon’s (1950) hierarchical model of intelligence also includes scientific and mathematic reasoning under a common factor ($k:m$).

In educational psychology the term quantitative reasoning is used more broadly to encompass both mathematical and scientific tasks (Halpern et al., 2007). Educators attempt to measure these two aspects in isolation through standardised tests of mathematical or science achievement, though one often involves drawing on content and skills from the other and performance on these tests are strongly correlated ($r > .80$; see Chapter 6). The issue of sex differences in mathematical performance has received greater attention than that of science achievement, but both domains are equally important for the problem of the underrepresentation of women in STEM. This is because developing a sense of mastery and competency in these domains is strongly
SEX AND SEX-ROLE DIFFERENCES IN COGNITIVE ABILITIES

linked to the decision to undertake further studies in STEM-fields (Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Simpkins, Davis-Kean, & Eccles, 2006).

Hyde, Fennema, and Lamon (1990) conducted the first meta-analysis of studies examining sex differences in mathematical performance. The results showed a weighted mean effect size favouring males of $d = +.20$. Hyde et al. then excluded from their effect size calculation all data from the Scholastic Aptitude Test of Mathematics (SAT-M), noting that the weighting would have inflated the overall effect size calculation, reaching a final effect size of $d = +.15$. The authors therefore concluded that there were negligible sex differences, with males scoring somewhat higher on some tasks and females scoring higher on others. However, this interpretation of the analysis has been challenged by other prominent researchers. Halpern and Wright (1996) noted that Hyde et al.’s conclusion obscures important differences across type of sample and task, as well as age-related trends. For example, in their meta-analysis, Hyde et al. found strong developmental differences across age groups. Young girls (age groups 5-10, and 11-14) scored slightly higher in mathematical performance than boys ($d = -.07$), but this pattern reverses shortly after puberty and upon entering high school as noted earlier by other researchers (Maccoby & Jacklin, 1974; Nash, 1979), with a stronger intensification of sex differences in the period between 8th and 12th grade. Students aged 15-18 show a small male advantage ($d = +.29$) in the high school years even after excluding the SAT-M, but the gender gap widens to a medium-sized difference in college years ($d = +.41$, 19-25) and beyond ($d = +.59$, 26 and older) which Halpern and Wright argued cannot be dismissed as being ‘trivial’. It also mirrors the developmental trend found for the Arithmetic subtest in intelligence tests, with no significant differences found in the
WISC but substantially higher scores in males under the WAIS for adolescents and adults (Kaufman & Lichtenberger, 2006).

Hyde et al. also found that the representativeness of the sample was a strong moderator – gender gaps were considerably larger in highly select samples (e.g., academically gifted, high SES), and lower in samples drawn from the general population. This raises the question of whether convenience samples demonstrate a selection bias, and whether sex differences in mathematics might be found with more representative samples. The joint effects of selection and variability is a methodological issue Becker and Hedges (1988) warned about when testing hypotheses about sex differences and similarities.

A subsequent meta-analysis by Lindberg, Hyde, Petersen, and Linn (2010) reached similar conclusions to their earlier review supporting the null hypothesis, but also included more representative nationally representative data from the NAEP and several longitudinal studies. Heterogeneity analyses of effect sizes showed that these studies were not comparable “and vary along some dimension(s)”, p. 1130. Moderator analysis presented by Lindberg, Hyde, Petersen and Linn found differences across nationalities, with U.S. and Australian samples showing a modest male advantage of $d = +.10$, and that there were age-related differences for high school students $d = +.23$ and college $d = .18$. However their conclusion was the same as earlier that it indicated no sex difference. A further limitation of their analysis was that they did not examine what the combined effect of mean sex differences and greater variability had on the sex-ratios of high achieving mathematical students, which earlier researchers identified as being considerably sized (Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000). Both studies
also constrained their focus to mathematical performance, and did not consider scientific reasoning which is typically larger (Halpern et al., 2007).

Given the issues of representativeness and other limitations noted above, a stronger evidence base is needed to test the hypothesis of sex differences in quantitative reasoning. Hedges and Nowell (1995) conducted a meta-analysis on sex differences in mathematics and science, drawing on large nationally representative assessments of U.S. student performance student testing data over three decades (1960-1992). Males significantly outperformed females across all years, with small effect sizes \( d = +.11 \) to \( d = +.26 \) found for mathematics and small to medium effect sizes for science \( d = +.11 \) to \( d = +.50 \). The authors also examined the combined effect of mean sex differences and greater male variability by examining the sex ratios of those performing at the upper-right tail of the ability distribution \( (95^{th} \text{ percentile}) \). High achieving males outnumbered females by a factor of 2:1 in mathematics, and ranged from 2.5:1 to as high as 7:1 for science achievement. While a pioneering study, the considerable passage of time since their analysis and purported changes in gender stereotypes and the status of women has led some authors (e.g., Caplan & Caplan, 2005) to question whether similar outcomes would be found with modern samples (an issue addressed in Chapter 4).

Another line of evidence on the magnitude of sex differences comes from the Scholastic Assessment (formerly, Aptitude) Test of Mathematics (SAT-M), which plays a critical effect on college admissions in the United States (see Figure 2.1). Halpern and Wright (1996, p. 6) noted that the gender gap in SAT-M scores has a ”huge” impact on college admissions and cannot be easily dismissed as being insignificant. A meta-analysis (see Appendix A1) on SAT-M results shows a weighted average effect size of \( d = +.30 \) \([95\%CI= .29 \text{ to } .31]\), drawn from a sample size of 30.3 million over the past
two decades (1996-2016). However Spelke (2005) has argued that inferring sex differences from such data may be problematic in that it is not a representative sample of the general population: the decision to sit the SAT-M exam is made by the student (and hence self-selected), and significantly more girls sit the exam than boys. The sample is, however, representative of potential college applicants and the sex difference effect has persisted over time with minimal change.

![Historical trend for Scholastic Assessment Test for Mathematics (SAT-M)](image)

*Figure 2.1.* Sex differences in mathematical performance on the SAT-M over the past two decades (1996-2016). Source: College Board Total Group Profile Reports.

Another useful source of information are standardization samples used in development and validation of tests of academic achievement. In order to develop tables of statistical norms, samples must be recruited across a range of ages that reflect the underlying population in terms of socioeconomic status, ethnicity, level of education, etc. These norms are used to evaluate a subject’s performance relative to other similarly
aged peers, for the purposes of evaluating giftedness as well as conditions such as dyscalculia.

Kaufman, Kaufman, Liu and Johnson (2009) reported data from the adult standardization sample of the *Kaufmann Test of Educational Achievement II, Brief Form* (KTEA-II Brief). The Math Composite subtest measures subjects on a broad range of arithmetic and applied problem-solving tasks. Adult males scored significantly higher than females, $d = .28$, which is a small but non-trivial effect size. Similarly, Camarata and Woodcock (2006) examined gender differences in the quantitative reasoning factor ($Gq$) for the *Woodcock-Johnson Achievement Tests* (WJ-77, WJ-R, WJ III), which comprises two mathematical subtests. Math Fluency measures the ability to add, subtract, and multiply rapidly and is a test of speed, while the Quantitative Concepts test involves mathematical formula and identifying number patterns. The authors found a small gender difference in favour of males across all ages $d = .16$, but which reached a peak for ages 19-34, $d = .47$ which is a medium effect size. More considerable gender differences existed for subjects aged 50-79, ($d = .84$) but this cannot be separated from the historical educational disadvantage experienced by women of that generation and are likely to be a cohort effect.

In the United States, an assessment comparable to the SAT-M is also conducted across various fields of science. High school students can elect to undertake rigorous coursework and complete the College Advanced Placement (AP) examinations, earning course credit in their chosen subject area and preparing them for college. Sex differences are frequently observed in college AP exams (Buck, Kostin, & Morgan, 2002; Moore, Combs, & Slate, 2012). For example, a meta-analysis of sex differences in performance on AP exams in the 2016 year\(^1\) showed small to medium sized effects in

---

\(^1\) Source College Board Advanced Placement Exam Dataset (https://research.collegeboard.org/programs/ap/data/archived/ap-2016)
Biology \( (d = +.27) \), Computer Science \( (d = +.10) \), Chemistry \( (d = +.31) \), Physics – Electricity and Magnetism \( (d = +.17) \), Physics – Mechanics \( (d = +.32) \), but no significant difference in Psychology \( (d = .00) \). Although these samples are also self-selected, they do indicate the presence of non-trivial sex differences in science achievement at the end of compulsory schooling in the United States. Willingham and Cole (1997) also report data from nationally representative samples of twelfth grade students, showing small to medium effect sizes across general tests of science (average \( d = +.17 \)). They also report data on the Armed Services Vocational Battery (ASVB) tests of general science \( (d = +.36) \), mechanical comprehension and reasoning \( (d = +.83) \) and understanding of electronics and circuits \( (d = +.78) \) which present considerably harder test content.

While these lines of evidence are highly US-centric, further conclusions can be drawn by examining cross-cultural patterns of sex differences in educational achievement. Guiso et al (2008) examined student testing data from the 2003 Programme for International Student Assessment (PISA), which offers a measure of student achievement in reading, mathematics and science in students at age 15 which is typically the final year of compulsory schooling in OECD nations. Their analysis found that boys scored higher than girls in mathematics in all but three of the 40 participating nations (Guiso et al., 2008, Figure S1A), but that the magnitude of the gender gap was diminished in countries with higher national levels of gender equality. Thus the presence of a gender gap in mathematics is not inevitable, and shows substantial cross-cultural variability. A meta-analysis by Else-Quest et al. (2010) on the same 2003 wave reached a similar conclusion, but also extended this to include younger participants from the 2003 Trends in Mathematics and Science Study (TIMSS). Though science achievement was also measured in PISA and TIMSS, the authors of both studies did not
report an analysis of this dataset (an issue addressed in Chapter 5). However a subsequent study by Reilly, Neumann and Andrews (2017)\(^2\) did examine mathematics and science achievement in younger students measured as part of the 2011 TIMSS wave. The study found substantial cross-cultural variability in size and direction of sex differences. Some countries reported higher male performance in science, while other countries showed a reversal of direction and reported higher female performance. Thus gender differences in science achievement, like mathematics, appear to be highly malleable to sociocultural factors.

Compounding the difficulty of interpreting whether sex differences in quantitative reasoning are meaningful is the test and grade discrepancy. When males as a group have higher test scores than females on tests of mathematics in science, it is presumed that this represents genuine differences in latent ability (i.e., actual quantitative reasoning skills). However, in an analysis of high school transcript data from the United States (where boys score higher than girls on NAEP mathematics and science tests), Shettle et al. (2007) observed that girls actually earn higher grades in mathematics and all fields of science. Subsequently, a meta-analysis by Voyer and Voyer (2014) found the same effect across a longer time frame with girls earning higher grades than boys in mathematics and science, at least for the United States. Dekhtyar et al. (2018) report a similar pattern of higher female grades in mathematics and science for students in Sweden. Thus one cannot preclude the possibility that (especially on high-stakes tests) girls show diminished performance due to factors such as test anxiety and stereotype threat, but are equal or greater in actual ability when alternate forms of assessment (such as laboratory reports, essays, and assignments) are employed.

\(^2\) Omitted due to space requirements
Alternately the sex differences observed for language tasks may be responsible for higher grades with alternate assessment to solely high-stakes exams.

While the magnitude of sex differences in mathematics and science may be disputed, one area where there is clear consensus among researchers is on attitudes and self-efficacy towards mathematics. In a meta-analysis Hyde et al. (1990) found that males report greater confidence and more positive attitudes towards mathematics across all age groups, and such findings are routinely reported in the literature. Insufficient studies have been undertaken for a science-focused meta-analysis, but Reilly, Neumann and Andrews (2017) report data from the 2011 wave of TIMSS, noting that boys in eighth grade reported more positive attitudes to science and greater science self-efficacy. Else-Quest et al. (2013) has argued that mathematics attitudes and self-confidence might be a stronger contributor to the underrepresentation of women in STEM than actual aptitude, though attitudes and abilities appear to be strongly intertwined. Nosek et al. (2009) also found that the strength of national implicit associations of mathematics with masculinity predicted the magnitude of sex differences in PISA test scores for mathematics and science achievement.

Collectively, these meta-analyses affirm the position held by Maccoby and Jacklin (1974) that sex differences in quantitative reasoning are well established, but that there may be a lack of consensus on the magnitude of the gender gap for modern samples. Additionally, gender stereotypes associating mathematics and science with masculinity and lower self-efficacy beliefs for females may be contributing factors.

**2.2.4 Memory.**

Evidence from a variety of studies suggest that there are observable sex differences in memory (Halpern, 2011), but Herlitz, Nilsson and Bäckman (1997) remark that the issue of sex differences in memory has been largely overlooked by
researchers in narrative and literature reviews. Stumpf (1995) notes that research in this area has been hampered by issues of reliability with instruments, representativeness of samples, and that performance differs across modality (e.g., verbal memory, spatial location, etc.). This has led to inconsistencies across studies, as typically only one type of memory is tested in a given sample. Additionally there are methodological issues with the level of difficulty of memory tasks. For example, recognition tasks (indicating whether the target stimuli matches that seen previously) are significantly easier than recall tasks (provide the name, type or location without retrieval cues) (Geffen, Moar, O'Hanlon, Clark, & Geffen, 1990). Additionally there exists considerable research on sex differences in memory using animal models (Jonasson, 2005), but this is not reviewed as research findings diverge considerably from patterns observed in humans.

In order to provide a comprehensive assessment of sex differences in the ability to memorise material, Stumpf and Jackson (1994) examined memory performance (across verbal and visual modalities) as part of a larger cognitive assessment battery used for the screening of applicants for medical school in West Germany. Their sample was large (approximately 97,000), and collected over a period of a decade. Women performed significantly higher than men on a common memory factor, with an effect size of \( d = -.56 \) (medium) across tasks.

Only a handful of other studies have tested memory across multiple modalities, often to examine the effects of aging in clinical or non-clinical samples of older persons (R. D. Hill et al., 1995). Unfortunately this limits the conclusions that can be drawn as they cannot be generalised to younger samples, and similar concerns may be raised about the representativeness of college-aged students from subject pools. But there is one other study that has investigated sex differences in memory using a community sample, and across multiple modalities. Herlitz et al. (1997) investigated differences in
short-term memory using a range of traditional and novel measures. Subjects were asked to memorise a series of 12-word lists while completing a distractor task (sorting playing cards), as a measure of verbal memory. They were also tested on their ability to recall newly acquired facts, by being asked to memorise a series of 20 fictitious statements about celebrities (e.g., “Person X collects stamps as a hobby”), and asked to later recall the statement associated with each person. Subjects also completed a variety of cued recall and recognition tasks for objects, and were asked to remember the name associated with a series of 16 colour photos of ten year old children as a measure of visual memory. Subjects were also asked at the end of the session to recall and name the activities they had participated in, testing episodic memory. Women scored significantly higher across all short term memory tasks, with small to medium effect sizes. But there was no sex difference reported on a control measure of general knowledge and vocabulary.

Most research on sex differences in memory focus on a single modality for tighter experimental control, with verbal learning memory being the most frequently tested. Typically this involves learning lists of words with a distractor task or time interval. For example Kramer, Delis and Daniel (1988) examined a community sample of 136 men and women matched on age and educational level. Subjects completed the California Verbal Learning Test, which requires subjects to memorise a list of 16 words across five trials. The number of words recalled is the dependent variable, and then subjects are given an interference list followed by a 20 minute delay. This allows for a delayed recall measurement. Women performed significantly higher than men, with a medium effect size across trials (from $d = -.37$ to $d = -.50$). A more detailed study by Geffen et al. (1990) differentiated between recall and recognition verbal memory. They recruited cross-sectional community sample of neurologically healthy adults aged 16-
86, reporting separate analyses for immediate recall \((d = -.57)\), twenty minute delayed recall \((d = -.48)\) and recognition \((d = -1.0)\) tasks. These studies are representative of verbal memory studies in methodology and findings, but to date a meta-analysis has not been conducted across studies. It is unclear whether sex differences in memorising verbal lists of words is the result of superior verbal and language abilities in females or instead the result of sex differences in episodic memory (Herlitz & Rehnman, 2008). Additionally, Lynn and Irwing (2008) reported a meta-analysis of performance on the digit span tasks of the WISC/WAIS, which requires subjects to store sequences of digits in memory and repeat them back to the examiner in the same order as they were presented. The effect size is considerably smaller \((d = -.13\) for girls, \(d = -.12\) for women), but consistent with the pattern of higher female performance for verbal memory tasks.

A number of studies have also investigated sex differences for visual memory. Patterns of sex difference differ across tests. Males tend to score slightly higher on visual-spatial working memory, with a meta-analysis reported by Voyer, Voyer and Saint-Aubin (2017) finding a small sex difference of \(d = +.15\) across studies. This might well be a result of the general male advantage on visual-spatial tasks. However it is equally possible that greater visual-spatial working memory affords an advantage when attempting visual-spatial problems. However, studies employing other types of visual memory tasks have found females score better than males, and these studies are often overlooked in reviews that conflate spatial working memory with visual memory more generally (e.g., A. C. Hill, Laird, & Robinson, 2014). For example girls and women were more adept at recognising introduced faces and recalling their names (Larrabee & Crook, 1993; West, Crook, & Barron, 1992). A meta-analysis by Herlitz and Lovén (2013) found a small to medium effect size across studies, \(d = -.36\) for learned faces.
Contrary to gender stereotypes about navigation, Galea and Kimura (1993) found that females are better able to memorise landmarks, learn routes on a map, and make fewer errors than males, with medium-sized effects $d = -0.45$.

Females are also more adept at remembering information presented visually. Silverman and Eals (1992) developed the *object identity memory task*, which presents subjects with an array of line drawings (such as common household objects, animals, tools, etc.) that they were asked to memorise. After a delay subjects were presented with a second page containing new and old items. Women were significantly better than men at recognising items that did not match the original, and this finding has been subsequently replicated with children and other adults. Another type of visual memory task that shows greater female performance is memory of object locations. Crook, Youngjohn and Larrabee (1990) first reported a sex difference for object locations after designing the Misplaced Objects Test, which presents subjects with a list of 20 household objects located on an overhead map of a simulated house. After a delay of 40 minutes they are asked to recall the locations of the objects, with females scoring significantly higher than males with a small to medium effect size $d = -0.36$. This and a similar study by Silverman and Eals sparked a wave of subsequent studies on visual memory for objects. A subsequent meta-analysis by Voyer (2007) found these effects to be robust, with a small female advantage for object identity $d = -0.23$, and object location tasks, $d = -0.27$. Collectively these results suggest higher female performance for most visual memory tasks, but that a small non-trivial male advantage exists for visual-spatial working memory.

Some studies have also examined whether there are sex differences in memory for identification of voices (sometimes termed earwitness recognition), but the findings are inconsistent. While several studies found superior female performance, especially
when hearing female voices (Roebuck & Wilding, 1993; Wilding & Cook, 2000), other studies found no meaningful difference (Yarmey, Yarmey, Yarmey, & Parliament, 2001). Finally one study examined sex differences for odour recognition, finding women were better at recognition of learned odours after study-test intervals of up to 21 days (Lehrner, 1993).

In summary, regardless of the modality chosen, women demonstrate superior episodic memory abilities (Herlitz et al., 1997; Herlitz & Rehnman, 2008). The sole exception to this general rule is for visual-spatial working memory (Voyer et al., 2017). Such studies generally find around a small to medium effect, though as Stumpf noted this is an under-researched area and there may be sex differences in other aspects of memory that have yet to be examined.
### Summary of Sex Differences in Specific Cognitive Abilities

<table>
<thead>
<tr>
<th>Class</th>
<th>Category</th>
<th>Developmental trend</th>
<th>Direction</th>
<th>Magnitude of Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Ability</td>
<td>Reading</td>
<td>Increases with age</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>Increases with age</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phonological Coding/Perceptual Speed</td>
<td>Unclear from literature</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>Small until late adolescence, may reach parity in adulthood</td>
<td>F &gt; M (child)</td>
<td>No diff. (adult)</td>
</tr>
<tr>
<td></td>
<td>Verbal Fluency</td>
<td>Increases with age</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Listening Comprehension</td>
<td>No effect</td>
<td>No diff.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spelling</td>
<td>Increases with age</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grammar</td>
<td>Increases with age</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speech Production</td>
<td>Increases with age</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>Category</td>
<td>Developmental trend</td>
<td>Direction</td>
<td>Magnitude of Effect</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------</td>
<td>------------------------------</td>
<td>-----------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Verbal Ability</td>
<td>Second language Acquisition</td>
<td>Unclear from literature</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td>Visual Spatial</td>
<td>Verbal Reasoning</td>
<td>Unclear from literature</td>
<td>M &gt; F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mental Rotation</td>
<td>Increases with age</td>
<td>M &gt; F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spatial Perception</td>
<td>Increases with age</td>
<td>M &gt; F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spatial Visualization</td>
<td>Increases with age</td>
<td>M &gt; F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spaciotemporal Ability</td>
<td>Unclear</td>
<td>M &gt; F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mathematical Reasoning</td>
<td>Increases with age</td>
<td>M &gt; F</td>
<td></td>
</tr>
</tbody>
</table>

Note: The direction of differences is indicated as M > F (Men > Women) or F > M (Women > Men).

The magnitude of effect is represented as:
- No difference: $d = 0$
- Small: $0.10$ to $0.20$
- Medium: $0.30$ to $0.40$
- Large: $0.50$ to $0.60$
- Extra large: $0.70$ to $0.80$
<table>
<thead>
<tr>
<th>Cognitive Domain</th>
<th>Performance</th>
<th>Age Effects</th>
<th>Sex Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative reasoning</td>
<td>Scientific Reasoning</td>
<td>Increases with age</td>
<td>M &gt; F</td>
</tr>
<tr>
<td>Memory</td>
<td>Verbal Memory</td>
<td>No age effects</td>
<td>F &gt; M</td>
</tr>
<tr>
<td></td>
<td>Visual Memory</td>
<td>Unclear</td>
<td>F &gt; M</td>
</tr>
<tr>
<td>Object identity/location</td>
<td>Increases with age</td>
<td></td>
<td>F &gt; M</td>
</tr>
<tr>
<td>Visual-Spatial WM</td>
<td>No age effects</td>
<td></td>
<td>M &gt; F</td>
</tr>
</tbody>
</table>

Note: Literature review of visual-spatial reasoning is deferred until Chapter 3.
2.2 Popular Beliefs about Intelligence

Intelligence is a socially desirable trait across most - if not all - cultures (Alicke, 1985; Shackelford, Schmitt, & Buss, 2005), and most lay persons have intuitive beliefs about what constitutes “intelligence” even when they lack the ability to formally define it. Ordinary laypersons may operate with somewhat different perceptions of what constitutes intelligence than are held by intelligence researchers. Furthermore, there are commonly held beliefs (“folk wisdom”) about sex differences in intelligence which hold that males and females hold different intellectual capabilities (Halpern, Straight, & Stephenson, 2011), that they have different “learning styles” and ought to be taught in separate classes (Halpern, Eliot, et al., 2011), or that hormones affect cognition to such a degree that one gender or another might be unsuitable for certain occupational tasks (Halpern, 1997). Indeed, I have personal experience of such folk wisdom. After having one of my earlier papers on sex differences in visual-spatial ability accepted for publication in a moderately prestigious feminist psychology journal, I learned that the publisher had (independently of the journal) prepared a press release highlighting that study as scientifically confirming the folk wisdom that women cannot drive or navigate as well as men. Needless to say this was not a desirable outcome and the press release was cancelled, but it does serve to illustrate the readiness in the community to confirm commonly held stereotypes about sex differences in intellectual functioning.

The beliefs of laypersons on the intersection of sex and intelligence are important to consider for several reasons. Firstly, they demonstrate in our culture a readiness to anticipate and confirm that there are inherent differences between the sexes in intellectual capabilities (even when there are not). Secondly, sex stereotypes held by parents, teachers, and the wider society may become self-fulfilling prophecies (Rosenthal & Jacobson, 1968), in that they subtly influence the way that children are
treated. Parental and teacher expectations are important socializers of sex stereotypes, exerting an effect at home and in the educational system (Hyde & Lindberg, 2007). Thirdly negative sex stereotypes may be internalised by individual boys and girls which leads to gendered differences in self-concept, self-efficacy, interest in scholastic subjects and motivation to achieve (Parker, Van Zanden, & Parker, 2018).

Halpern (1997, p. 1091) has claimed that the question of sex differences in intelligence “is among the most politically volatile topics in contemporary psychology”. This argument was made because of the polarising effect the topic can have on laypeople, and the implications that it can hold for the way in which boys and girls are educated. Hare-Mustin and Marecek (1988) conceptualised these two divergent positions as alpha bias (assuming large and immutable sex differences, or the “Men are from Mars, Women are from Venus” view popularised by Gray) and beta bias (assuming minimal or non-existent differences between males and females, such as the position held by Hyde and colleagues). The two perspectives are also worth considering when critically evaluating research in this area, as the ideological position of individual researchers may be relevant in cases where a highly selective review of the literature is presented that omits key evidence (e.g., Lippa, 2006 in critiquing Hyde (2005)).

### 2.1.1 Self-Estimation of Intelligence.

The way in which we see ourselves can profoundly impact our educational aspirations, as well as our academic self-esteem and self-efficacy. So an important theoretical question is just how accurate are our self-evaluations of intelligence? Kruger and Dunning (1999) espoused the rather pessimistic view that people generally hold overly optimistic views of their own intellectual functioning, which has been termed the “above-average effect” (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Kruger, 1999), also is sometimes referred to as the “Lake Wobegon” effect after the
fictional American town. Essentially this is the tendency of the typical person to believe she or he is somewhat above average in intelligence (which for a normally distributed psychological characteristic defies statistics). Upholding a self-concept that is more positive than objective reality may actually be adaptive and serve a number of psychological functions (e.g., esteem maintenance) and translate to improved self-efficacy (Alicke, 1985; Greenwald, 1980). While the above-average effect has been consistently replicated, it is also important to note that not every person exhibits unrealistically high estimates. Under some conditions a related “below-average-effect” can be found, particularly when social comparisons are made to groups involving positive stereotypes (Kruger, 1999), such as stereotypes about racial grouping or gender.

Despite the cognitive bias of the above-average effect, when asked to provide an estimate of their own intelligence relative to others, males consistently report higher estimates than do females. This effect was first reported by Hogan (1978) who found that females provided lower estimates of their IQ score than males in an American sample, and was subsequently replicated with a British sample (Beloff, 1992). Although actual psychometric intelligence was not measured in these studies, given that actual IQ does not differ between the sexes in the general population these seemed surprising findings with no definitive cause. The researchers speculated that they reflected either cultural sex stereotypes, or a strategy of self-minimisation on the part of females for social desirability reasons and boastfulness for males.

These studies sparked considerable research into sex differences in self-estimated intelligence (SEI), defined as people’s estimates of their own intellectual abilities relative to the general population. This is typically assessed by presenting participants with a histogram bell curve with appropriate text anchors, and instructions
explaining the bell-curve distribution of intelligence in the general population. The problem has been termed by Furnham et al. (2001) as the *male-hubris female humility* problem (MHFH), and has been replicated cross-culturally across dozens of nationalities. While initial studies focused exclusively on overall IQ, research has been extended to cover more specific cognitive abilities. In a meta-analytic review, Syzmanowicz and Furnham (2011) found robust sex differences for general intelligence ($d = .37$), mathematical intelligence ($d = .44$), visual-spatial intelligence ($d = .43$). Somewhat surprisingly there was also a marginal but significant effect for estimates of verbal intelligence ($d = +.07$), which runs contrary to the general research finding of greater verbal abilities in females. While the effect itself seems robust, sex differences in SEI are not easily explained and may reflect a combination of social motives (boastful hubris on the part of males, humility on the part of females), courtship strategy, sex stereotypes about intellectual ability, or the pattern of lower general self-esteem in females (Gentile et al., 2009; Kling, Hyde, Showers, & Buswell, 1999). The latter explanation remains as yet untested, but is investigated in Chapter 10.

**2.1.2 Estimation of other’s intelligence.**

If social motives or courtship strategy were the sole mechanisms involved in SEI, then presumably when asked to provide an objective estimate of other people’s intelligence the hubris/humility effect should not be present. In the original study by Hogan (1978), participants were also asked to provide an estimate of the intelligence of their mother and their father. Fathers were rated as more intelligent than mothers, even though there are no sex differences in general intelligence present in the community. The effect has been replicated numerous times (Beloff, 1992; Furnham & Rawles, 1995), but such an effect should be interpreted cautiously. It could be argued that the effect actually be a reflection of systemic educational and occupational inequalities of
the time (i.e., higher male educational advancement) rather than genuine beliefs that men are inherently “smarter”.

Furnham and Gasson (1998) took a different approach, and instead asked parents to provide an estimation of the intelligence of their own children. Parental beliefs may be a particularly important mechanism in the socialisation of sex stereotypes, as parental educational expectations may influence a child’s own view of their capabilities (Jodl, Michael, Malanchuk, Eccles, & Sameroff, 2001). Sons were rated as more intelligent than daughters \( (d = .67) \), and this effect has also been replicated several times (Furnham, 2000; Furnham, Reeves, & Budhani, 2002). Thus such a pattern might better fit with cultural beliefs about gender stereotypes.

2.1.3 Popular beliefs about sex differences in specific cognitive abilities.

As mentioned above, one apparent explanation for sex differences in SEI might be the effect of gender stereotypes (e.g., males are inherently “smarter”, and hence their abilities are over-estimated). Surprisingly few studies have actually sought to test this hypothesis, and determine whether laypersons are accurate judges of psychological sex differences. Swim (1994) found across a range of behavioural characteristics, laypersons were consistent in the direction of sex differences but tended to underestimate their magnitude (beta bias). However Swim did not include any specific questions on intelligence differences, rather focusing on a subset of cognitive abilities. More recently Halpern, Straight and Stephenson (2011) specifically asked for estimations about sex differences in cognitive abilities to determine their accuracy. They asked participants to estimate cognitive sex differences across a range of 12 tasks, and then compared the estimates provided to actual sex difference studies. Their results showed that participants were generally accurate about the direction of sex differences for specific cognitive abilities, but underestimated their magnitude. Crucially though,
their study did not enquire about *global* differences in intelligence (IQ). If there are gender stereotypes about overall intelligence, it would appear that these are implicit biases (expressed in SEI or about other individuals) rather than explicitly held beliefs about a smarter sex (i.e., males/females are generally smarter).

One notable exception to beliefs about sex differences in intelligence appears to be the ability of cognitive multitasking (defined as the ability to execute two or more cognitive tasks concurrently, or using task-switching). Szameitat et al. (2015) reported the results of large study recruiting from the UK, USA, Germany, Netherlands and Turkey, finding that over half of study participants expressed a belief in sex differences in multitasking. Of those participants, 80% believed that women were simply better multi-taskers, despite a lack of research studies to support such a belief (Szameitat et al., 2015). Thus not all lay gender stereotypes reflect negatively on women, and some may suggest women have a modest advantage in some skills even if not empirically supported.

### 2.3 Theoretical perspectives on sex differences in cognitive abilities

While sex differences in cognitive ability have been studied since the beginning of psychometric assessment of intelligence (for a review see Shields, 1982), theoretical perspectives on their origins have shifted considerably over this time period. Early theoretical debate on sex differences proposed strong and immutable biological factors (*nature*), while later theorists argued that early differences in socialisation experiences of boys and girls and environmental factors (such as parental and teacher expectations, gender stereotypes, etc.) might better explain differences in cognitive ability (*nurture*). More recently, the limitations of both the nature and nurture perspective in isolation have been recognised (Halpern & Tan, 2001; Priess & Hyde, 2010). Increasingly, sex difference researchers acknowledge the need for a more comprehensive theoretical
framework that encompasses nature, nurture, which Halpern and Tan (2001) has termed a psychobiosocial model of sex differences. Less seldom acknowledged by researchers are macro-level sociocultural factors contributing to male-female group differences. These include cultural beliefs and practices (such as gender stereotyping of intellectual domains like language and STEM), as well as larger cultural issues such as general attitudes towards women, gender equality, and the stratification of educational, occupational and political representation along gendered lines (termed the gender segregation hypothesis). These macro-level factors may help to explain some of the cross-cultural variability in the magnitude of sex differences in cognitive ability.

### 2.3.1 Biological Explanations for Sex Differences

Men and women differ in a number of important characteristics beyond their reproductive capabilities, including height, physical strength, and lifespan. Any explanation for sex differences in cognition must rightly consider whether these make a contribution. But in doing so, one must also evaluate such claims with a stronger degree of intellectual rigour than was present in earlier centuries. Take for example the observation that due to their natural height advantage, males on average have slightly larger brains than females - but recall that also men and women do not differ significantly on tests of general intelligence (psychometrically measured IQ). The assumption had been made by early anatomists that larger brains would confer some degree of intellectual advantage (for an historical account see Shields, 1975), and was mounted as a scientific argument for the intellectual infirmity of women. Indeed, there are some current researchers (i.e. Lynn, 1994, 2016) who continue to make this argument (intelligence is correlated with brain size; males have larger brains; males must therefore be smarter and IQ tests showing no difference are wrong).
Advances in scientific methodology, such as neuroimaging, have meant that biological arguments have been held to greater scrutiny, and some once ‘well established’ biological claims have been found to stand on shaky foundations. For example, support for putative sex differences in brain structure is now highly disputed (for a discussion see Fine, 2010b), and limited to a small number of brain regions including the hypothalamus, amygdala, and corpus callosum (Hines, 2010), as well as proportionately greater development of Broca’s and Wernicke’s area in females which are involved in speech production and language (Harasty, Double, Halliday, Kril, & McRitchie, 1997; Robichon, Giraud, Berbon, & Habib, 1999). Other claims have held up better, such as sex differences in lateralization of brain function for language (Levy, 1969; B. A. Shaywitz et al., 1995). Males who are right-handed tend to exhibit greater activity in the left hemisphere when performing language processing tasks, whereas females show a greater likelihood for bilateral pattern of brain activation (reviewed in Kansaku & Kitazawa, 2001). But even these findings are disputed (Wallentin, 2009), due to inconsistencies across studies (Sommer, Aleman, Bouma, & Kahn, 2004).

Although there has been a shift in the literature away from biologically-based arguments for sex differences in cognition, there remain several areas which maintain strong scientific support. These are reviewed herein, but the interested reader is cautioned that this list is not exhaustive. It is acknowledged that there are other theoretical perspectives (such as neuroanatomical differences in brain structure, or lateralization of brain function) that have been proposed, and are contentious. Fine (2010a, 2010b) summarises the debate on methodological limitations of neuroimaging studies, including extremely small sample sizes, confounding variables and failure to replicate findings. Attention has been given to those topics where there is some degree of scientific consensus. Three sources of evidence are outlined: effects of sex hormones
during early development, activational effects of sex hormones, and finally biological differences between males and females predicted by evolutionary psychology.

2.3.1.1 Sex hormones during early foetal development.

One of the most prominent explanations for sex differences in cognitive performance has been the effect of sex hormones (Kimura, 2000). While males and females produce both classes of sex hormones to some extent, typically there is greater androgen production in males and greater estrogen and progesterone production in females. This difference starts early, with observable differences in testosterone concentration of foetuses found as early as 8 weeks gestation (Hines, 2010). Production of androgens in males remains high between 8 and 24 weeks, then reduces shortly before birth (Hines, 2010). During this developmental window sex hormones contribute to the organisation and development of the brain, and shape gender development (Berenbaum & Beltz, 2011, 2016).

While sex-specific differences in hormone production outlined above are present, during some pregnancies the developing foetus will receive additional exposure to sex hormones. This exposure can affect both sexes (Auyeung et al., 2009) and may lead to enduring changes in brain development and sex-typed behaviour (Berenbaum & Beltz, 2016; Hines, 2015). For example, girls diagnosed with the condition congenital adrenal hyperplasia (CAH) are exposed to high levels of androgens prenatally, and exhibit a stronger preference for playing with boys’ toys than female relatives (Berenbaum & Hines, 1992; Hines, 2006). Such girls also perform better than their same-sex peers on tasks of spatial ability later in life (Puts, McDaniel, Jordan, & Breedlove, 2008). Longitudinal studies have also found that levels of foetal testosterone assayed from amniotic fluid is also associated with similar male-typical patterns of play behaviour and interests in boys and girls (Auyeung et al., 2009).
One indicator of early prenatal androgen exposure is the ratio between the length of the second (index) and fourth fingers (2D:4D) in adults. Some studies have found associations between lower 2D:4D ratios and performance on visual-spatial tasks for both sexes (Collaer, Reimers, & Manning, 2007; Falter, Arroyo, & Davis, 2006; Peters, Manning, & Reimers, 2007), but other studies have failed to find an association (Puts et al., 2008; Voracek, Pietschnig, Nader, & Stieger, 2011). The inconsistent evidence may be due to measurement issues associated with digit ratios, or that digit ratios are only a crude proxy for actual levels of androgen exposure (Miller & Halpern, 2013). When combined with the stronger evidence found in amniotic fluid assays and CAH studies (Auyeung et al., 2009; Hines, 2015), there appears to be some effect of sex hormones on the developing brain, even if it is only subtle.

The exact mechanism responsible for influencing cognition though is unclear – while prenatal exposure to sex hormones might have a direct effect on cognition, since such children also show increased male-typical behaviour and interests, they may be self-selecting situations that promote spatial development. Because spatial ability requires environmental input and practice for development (Baenninger & Newcombe, 1989), childrens’ toys and play can be important sources of spatial experiences and training. Many stereotypically masculine activities (such as construction blocks and models) promote spatial development (Doyle, Voyer, & Cherney, 2012), and in this way the developmental effects of sex hormones may be acting indirectly.

### 2.3.1.2 Activational effects of sex hormones.

While sex hormones contribute to brain development, they also exert an activational role on human cognition and behaviour. In normal development, the production of sex hormones increases considerably with puberty, and continues long into adulthood. This developmental milestone coincides with an intensification of
gendered interests and behaviours (Halim & Ruble, 2010), as well as a widening of the gender gap for many types of cognitive ability (Hyde, Fennema, & Lamon, 1990; Voyer et al., 1995), leading to the intuitive appeal of sex hormones as an explanation. However, research findings examining the association between endogenous sex hormone levels and actual cognitive tasks in men and women have produced only mixed support for such claims. Some studies have found little to no association (Halari et al., 2005; Puts et al., 2010), while other studies have found testosterone levels are correlated with visual-spatial performance and estrogen levels to performance on verbal tasks (Griksiene & Ruksenas, 2011; Hausmann, Slabbekoorn, Van Goozen, Cohen-Kettenis, & Gunturkun, 2000; Kimura, 1996). This suggests that the activational effects of sex hormones may indeed be quite subtle, and dependent on context. Some theorists have speculated that the effect of hormones may be mediated by other factors such as the expression of sex-roles (Smith, Deady, Sharp, & Al-Dujaili, 2013) and gender stereotypes (Hausmann, Schoofs, Rosenthal, & Jordan, 2009). Alternatively, the activational effects of sex hormones may be dwarfed by the earlier influences of hormones on brain development.

Further evidence for a contribution of sex hormones on cognition comes from samples that, due either to natural aging or specific medical conditions, are somewhat lower in natural hormone production. Production of testosterone decreases in men due to natural aging, but when given testosterone therapy they show improved performance on tests of spatial ability (Janowsky, Oviatt, & Orwoll, 1994). Women receiving hormone replacement therapy for the treatment of menopause show an improvement in their verbal memory (Maki, 2001) and verbal fluency (Maki, Rich, & Shayna Rosenbaum, 2002), but it is also associated with a decline in spatial ability (Drake et al., 2000).
2.3.1.3 Evolutionary psychological arguments of innate differences.

Another purported biological contribution to the emergence of sex differences in cognitive ability at the population level comes from the field of evolutionary psychology. Darwin (1871) first proposed that sexual selection as a result of evolutionary pressures led to a differentiation in the roles of men and women. In our distant hunter-gatherer past, men would be required to travel long distances to track and hunt animals, a task requiring strong spatial perception and navigation skills (Silverman & Eals, 1992). Those males who lacked an aptitude in spatial ability would be unable to accrue sufficient resources to thrive in traditional hunter-gatherer societies, and thus be less likely to propagate their genes to future generations (Buss, 1995). In contrast, women fulfilled the role of the gatherer of local food and assumed childrearing duties. Since these roles have less need for spatial proficiency but emphasize other adaptive traits such as nurturing and fine-motor skills (Halpern, 2000), over successive generations evolutionary forces may have developed sex-specific proficiencies. This argument has been expanded upon by evolutionary psychologists (Archer, 1996). Given that spatial ability is believed to lay down a foundation for quantitative reasoning, Geary (1996) proposed that evolutionary pressures led to group differences between males and females in spatial and mathematical reasoning. A limitation of this argument is that it is a post-hoc explanation for observed group differences, formulated to explain existing behaviour (Cornell, 1997). Furthermore, evolutionary psychology has yet to lay down a similarly compelling argument as to why sex differences in language and verbal abilities might emerge, especially for relatively “recent” traits (at least on an evolutionary time-scale) such as reading and writing proficiency.
2.3.2 Psychosocial Explanations for Sex Differences

The prominence of biological contributions to sex differences has declined in recent years, with an increasing recognition by researchers of the role of socialisation factors on psychological sex differences in behaviour and cognition (Leaper & Friedman, 2007). While there are indeed many similarities and shared experiences, from infancy onwards parents and caregivers treat boys and girls differently in sometimes subtle, and sometimes quite overt ways that contribute to a child’s gender development (Bussey & Bandura, 1999). A large number of psychosocial explanations have been proposed for observed sex differences in cognitive ability, including behavioural conditioning and modelling (social cognitive theory), cognitive perspectives such as gender schema theory (Bem, 1981b), gender stereotyping about the abilities of males and females, and so on. While an exhaustive list is beyond the scope of this review, the most prominent and recurring themes will be highlighted. Most are complementary rather than contradictory in that they acknowledge the contributions of other perspectives (Leaper & Friedman, 2007), but place greater emphasis on particular processes (C. L. Martin, Ruble, & Szkrybalo, 2002). All share a commonality in that they are proposed to arise from the different socialisation experiences of boys and girls.

2.3.2.1 Differential socialisation experiences.

One of the most frequently encountered responses parents receive when announcing a pregnancy is to be asked “Is it a boy, or a girl?”. While either of these options is a cause for celebration, it does serve to illustrate just how salient the attribute of gender is in our society. Parents and caregivers treat boys and girls differently in sometimes subtle, and sometimes overt ways that contribute to a child’s gender development (Bussey & Bandura, 1999; Siegal, 1987). For example, in an extensive meta-analysis of studies involving observations of parental interaction with their
children, Leaper, Anderson and Sanders (1998) found that mothers tend to be more talkative with daughters than sons, and use more supportive speech. Parents also tend to discuss emotional experiences with their daughters more than their sons, as well as using a wider and more nuanced range of emotional vocabulary (Adams, Kuebli, Boyle, & Fivush, 1995; Fivush, 2014; Flannagan & Perese, 1998). These varied experiences may provide girls and boys differential opportunities to practice and develop language proficiency (Clearfield & Nelson, 2006), as well as contributing to sex differences in emotional sensitivity and expression (Brody & Hall, 2000; Chaplin, Cole, & Zahn-Waxler, 2005). While boys and girls typically receive the same literacy instruction upon entering formal schooling, early childhood experiences that encourage language and scaffold literacy acquisition can be beneficial (Neumann, Hood, & Neumann, 2009; Neumann & Neumann, 2010). Some studies have found parental differences in teaching experiences with sons and daughters (Flannagan & Baker-Ward, 1996; Flannagan & Perese, 1998), but others find no child-gender effect (Tenenbaum & Leaper, 1998), or that child-gender effects may vary between mothers and fathers (Leaper, 2002). Further research is required to determine the extent of child-gender effects in parent-child conversations and teaching experiences, but there is tentative evidence to suggest some socialisation differences between boys and girls.

To investigate the proposition that daughters and sons are treated differently by parents, Lytton and Romney (1991) conducted a meta-analytic review of studies observing parent-child interactions across a number of socialization areas. Somewhat surprisingly, they found only weak evidence of differential socialization practices between boys and girls in many areas. It should be acknowledged, however, that their meta-analysis was limited to the empirical studies available at the time (for a review of more recent studies, see Leaper & Friedman, 2007). No doubt there are many
similarities and shared experiences of girls and boys, but there are also fundamental
differences that sustain the gender socialization process. One area in which this is
particularly apparent is in the encouragement (or discouragement) of sex-typed play and
activities. Lytton and Romney found that both mothers ($d = .34$) and fathers ($d = .49$)
strongly encouraged sex-typed activities and play in their children along traditional
gender norms. But children also actively contribute to this process as well. Children as
young as three years of age develop a preference for playing with children of the same-
sex (Maccoby, 1998; C. L. Martin & Ruble, 2004). Gender-atypical patterns of play
may be highly discouraged or even a source of ostracism (Egan & Perry, 2001; Young
& Sweeting, 2004). Recall also that children often express a preference for sex-typed
toys and play activities that is partially biologically driven (Alexander & Hines, 2002;
Hines & Alexander, 2008). Children may also self-select those that appear more
enjoyable or hold greater interest (particularly with increasing age and greater
autonomy) These choices may in turn be reinforced (or discouraged) by parents (Idle,
Wood, & Desmarais, 1993), and so the process may represent an iterative process
guided by a variety of biological and socialisation factors.

Piaget (1951, 1968) argued that play provides children with opportunities to
practice socially valued skills that contribute to their social and cognitive development,
especially when playing with parents and their peers. The nature of play has important
implications for cognitive development (for a review see Liben, Schroeder, Boriello, &
Weisgram, 2018). These early experiences also communicate messages to children about
sex-typed behaviour, and gender stereotypes. Two proposed mechanisms have been
suggested to contribute to greater sex differences in cognitive ability. Firstly, a number
of researchers had found associations between stereotypically masculine play and toys
and the development of spatial ability (Connor & Serbin, 1977; Sherman, 1967), which
has been suggested contributes to proficiency in quantitative reasoning. Baenninger and Newcombe (1995) showed that environmental input and practice is required to fully develop spatial ability, and a number of researchers have shown that gender stereotyping of activity and leisure interests predicts future spatial performance (Cherney & Voyer, 2010; Doyle et al., 2012; Signorella, Jamison, & Krupa, 1989). Differential opportunities to practice and develop their visual-spatial reasoning is therefore one mechanism by which the pronounced sex differences in visual-spatial abilities might emerge.

Secondly, differential socialization practices serve to reinforce stereotypes about the roles of men and women and divisions of labour in society (a point elaborated in Section 2.3.3.1 “Social Role Theory”). Cognitive skills that are seen as less relevant and less desirable for one sex may be less practiced, and hold less interest/motivation particularly in the face of difficulties (such as that experienced when reading for the first time, or solving difficult mathematics problems). For example, reading is stereotypically regarded as being feminine (Dwyer, 1974; McGeown, Goodwin, Henderson, & Wright, 2011; Steffens & Jelenec, 2011), and boys show reduced interest in reading than girls (L. Baker & Wigfield, 1999; Marinak & Gambrell, 2010; Millard, 1997). Conversely, mathematics and science are stereotypically associated with males and masculinity (Nosek, Banaji, & Greenwald, 2002; Steffens & Jelenec, 2011), and girls rate their mathematics and science ability considerably lower than boys (Eccles, Wigfield, Harold, & Blumenfeld, 1993; Reilly et al., 2017). Thus there may be multiple pathways whereby cultural gender stereotypes and socialization differences between boys and girls contribute to the emergence of sex differences in cognitive ability.
2.3.2.2 Cognitive-developmental model of sex-typing.

Kohlberg (1966) proposed a cognitive-developmental model of gender, emphasising that it was an active intellectual process rather than merely passive learning through behavioural reinforcement and punishment. Kohlberg was the first to move beyond observation of child behaviour towards their use of language, and to use interview methods to study childrens’ thoughts and beliefs about gender. Children quickly acquire their concepts about gender through a process known as sex-typing (Kagan, 1964; Kohlberg, 1966), the beginnings of which can be seen in the first two years of life (Kohlberg & Ullian, 1974; Perry, White, & Perry, 1984) and increases steadily with age. By age 3, most children can label themselves as a boy or a girl, and can label the gender of others with partial accuracy (Kohlberg & Ullian, 1974). Children develop an internal mental model of cultural prescriptions of maleness and femaleness which are termed sex-roles (for the roles and behaviours which were stereotypically associated with masculinity and femininity, but these expand with age to also include sex-specific self-concepts, personality traits, interests and cultural/intellectual pursuits). Kohlberg and Ullian (1974) argued that gender takes on a “tremendous importance in organizing the child’s social perceptions and actions” (p. 210)…. in essence, that gender takes on a centrality in a child’s perceptions of self and others, which Bem (1993) termed the “lenses of gender” for the way in which it colours our perspectives.

Kohlberg and Ullian noted that between ages 6 and 7, most children attain gender constancy (the belief that gender is fixed, and a boy cannot become a girl or vice versa), and begin to develop more sophisticated sex-role concepts about the sex-typing of toys, clothing, household objects and even occupations. Serbin, Powlishta, Gulko, Martin and Lockheed (1993) report that even by age 5 children attain an understanding of sex-typed personality traits, and can identify basic traits (such as “emotional”, or “strong”) that are stereotypically feminine and masculine. This finding has been
reproduced cross-culturally, and by age 8 most children show adult-equivalent performance on labelling tasks (Best, 1982; Best et al., 1977). Beliefs in inherent differences between males and females can be found across most cultures and time periods (Kite, Deaux, & Haines, 2008; Williams & Best, 1990), though the strength of these stereotypes does vary.

Kohlberg’s cognitive developmental model of sex-typing has held up well given the passage of time, and has laid down the foundations for other theoretical perspectives such as gender schema theory, and the role of gender stereotyping of activities and interests. It documented the process of sex-typing that other theories would incorporate, but makes no specific predictions about sex differences in cognitive abilities.

### 2.3.2.3 Gender Schema Theory and sex-role identification

An extension of Kohlberg’s cognitive-developmental model of gender was advanced by Bem (1981a), who proposed *gender schema theory* to account for the sex-typing process. The term cognitive schema denotes a cognitive structure or pattern of thought that helps organize and process categories of information and the relationships between them (Neisser, 1976). A schema helps guide our perceptions of the world, and helps draw our attention to information that is schema-relevant. Initially, children start out with a fairly primitive gender schema, noticing obvious physical differences such as height, clothing, and hair. The gender schema is later refined with more nuanced properties associated with gender such as behaviour, personality traits, and interests, in the developmental progression described by Kohlberg and later researchers (Halim & Ruble, 2010; Kohlberg & Ullian, 1974; Ruble, Martin, & Berenbaum, 2006). The gender schema then becomes a prescriptive standard to uphold, and motivates the individual to regulate his or her behaviour to conform to cultural definitions of masculinity or femininity (Bem, 1981a).
While an understanding of sex-role concepts is attained in early to middle childhood, there is wide variability in the degree to which individuals integrate aspects of femininity and masculinity into their personality. Earlier personality theorists had believed that masculinity and femininity represented a single personality trait (MF) that existed along a continuum as bipolar opposites (i.e., being high in one meant being low in the other). But in a landmark psychometric review, Constantinople (1973) presented a growing evidence base that they were not unidimensional and instead represented two distinct personality clusters (masculine and feminine), which Bakan (1966) labelled as agency and communality respectively. Men are thought to be agentic – assertive, competitive, achievement-driven, and dominant. By way of contrast, women are seen as more communal and expressive – friendly, cooperative, concerned with the wellbeing of others, and emotionally/verbally expressive. But sex-role identification does not always fall neatly into two binary categories of male and female. Some children identify with stereotypically masculine sex-roles (instrumental or agentic) while others identify more strongly with stereotypically feminine (expressive or communal) sex-roles. Still others acquire a combined blend of both masculinity and femininity into their self-concept, termed psychological androgyny (Bem, 1974; Spence & Helmreich, 1979). Less commonly, a small percentage of the population identify with neither masculine or feminine roles.

Sex-roles act as a self-regulatory mechanism: highly sex-typed children and adults are motivated to keep their behaviour and self-concepts consistent with the traditional gender norms of their biological sex (Maccoby, 1990; C. L. Martin & Ruble, 2004), which Egan and Perry (2001) have termed felt pressure. Such individuals are limited in their behavioural repertoire and will self-select activities and interests that conform to traditional gender norms. Others whose gender schema is more “fuzzy” and
less prescriptively defined may see masculine and feminine sex-roles as complementary, and incorporate both into their self-concept (Bem, 1983, 1984). Although felt pressure is usually considered in the context of sex-typed social behaviour, it also applies to performance in sex-typed academic domains such as language and mathematics/science (McGeown et al., 2011; Nosek et al., 2002; Nosek et al., 2009; Oswald, 2008; Steffens & Jelenec, 2011). A number of theorists (e.g., Nash, 1979; Sherman, 1967) have proposed that sex-role identification might be an important mediator of sex-differences in cognitive ability, but an elaboration on these issues and specific mechanisms is deferred until Section 2.3.4.

**2.3.2.4 Social Cognitive Theory.**

Bussey and Bandura (1999) proposed social cognitive theory as an explanation for sex differences. Social cognitive theory argues that childrens’ behaviours are shaped by reinforcements and punishments (coming initially from parents and caregivers, then transitioning to peers and other adults) as well as through the powerful influence of imitation and modelling of the behaviour of others (Bussey & Bandura, 1984). For most children, these learning experiences become internalised and form a core part of a child’s identity which both guides and restricts future behaviours, including gendered behaviour. This gendered behaviour can be expressed in various ways, such as leisure activities and interests, and provides the opportunity to practise and refine socially useful skills (such as language competence, or refining spatial ability). In this context, modelling may play a particularly important role, as young children develop aspirations to achieve a certain profession. Parents and teachers may also subtly transfer cultural stereotypes, particularly when certain areas of study are stereotypically perceived as masculine (like mathematics and science) or feminine (reading, languages, and social sciences).
2.3.2.6 Sex and gender stereotypes about cognitive ability.

Another theoretical perspective that has been proposed as contributing to sex differences in cognitive ability is that of sex stereotypes (now more commonly referred to as gender stereotypes, a practice adopted hereon). Beliefs that males and females possess different attributes and behaviours extends across all age groups, time periods and cultures (Kite et al., 2008), though the strength of such beliefs does vary across cultures and individuals. Gender schema theory cautioned that cultural prescriptions of masculinity and femininity can become self-fulfilling prophecies, a view shared by other researchers.

As outlined earlier, gender stereotypes are learned early and are highly resistant to change. Collectively, parents and teachers hold different educational expectations for boys and girls (Eccles, Jacobs, & Harold, 1990; Frome & Eccles, 1998; Jussim & Harber, 2005), especially for mathematics and science. Even in the absence of the overt endorsement of gender stereotypes, implicit stereotypes associating mathematics/science with masculinity and arts and language with femininity are strong (Nosek et al., 2002). Consequentially young girls in elementary school report lower self-efficacy and competence beliefs in mathematics than boys (Eccles et al., 1993), even though sex differences on standardised tests of mathematics achievement show minimal sex differences in this age group (Hyde, Fennema, & Lamon, 1990). Girls also report lower self-efficacy beliefs about science, and this pattern is found cross-culturally in adolescents (Reilly et al., 2017). Boys also report lower self-efficacy and competence in reading and language domains (L. Baker & Wigfield, 1999; Eccles et al., 1989; Eccles et al., 1993), although sex differences in actual reading ability can be quite large. This makes it difficult to establish whether lower self-efficacy and interest for reading are the product of negative gender stereotypes or a reflection of accurate self-appraisal.
Gender stereotypes are important to consider for several reasons. Firstly, reduced self-efficacy beliefs can be associated with avoidance behaviour and learned helplessness. If one believes that one’s ability is fixed and associated with factors outside of one’s control, motivation to improve will consequentially be lower. For example, when children are asked to complete a novel task and are told that the task is better suited for one gender or the other, those in the stereotyped group are less likely to persist when encountering difficulty and generally show impaired performance (Cimpian, Mu, & Erickson, 2012).

The two most prominent gender stereotypes are the association of reading and language proficiency as being feminine, and the association of mathematics and science as being inherently masculine. This can be regarded as a distal factor contributing to sex differences in cognitive ability, in that these gender stereotypes are learned early and shape how an individual approaches intellectual domains over the course of a lifetime. Secondly, there is evidence to suggest that even when one is quite proficient at a cognitive task, knowledge of group differences in ability can undermine cognitive performance through the mechanism of stereotype threat (Steele, 1997, 1998), especially for high-stakes standardised tests. As such, stereotype threat represents a more proximal factor contributing to sex differences in cognitive ability. It also raises the uncomfortable possibility that some (as yet unmeasurable) component of sex differences in cognitive ability may be an artefact of the testing environment and not an accurate reflection of actual ability. By way of example, a consistent finding in the literature is that girls and women achieve higher grades in most academic subjects, including mathematics and science subjects (Alon & Gelbgiser, 2011; Voyer & Voyer, 2014). Yet males score higher than females on standardised tests of mathematics and science (Hedges & Nowell, 1995; Reilly, Neumann, & Andrews, 2015), suggesting that
there is a gap between performance in testing situations and achievement overall (Halpern, 2011). Any comprehensive psychobiosocial model of sex differences must consider the role of gender stereotypes (both distal and proximal) in the emergence of sex differences in cognitive ability at the population level (Halpern & Lamay, 2000).

An important example of the effect of self-fulfilling prophecies in an educational context comes from the classic educational study by Rosenthal and Jacobson (1968). The researchers had children complete a purported screening test for intellectual ability in first grade, and identified to teachers a subset of children as intellectually gifted for their age. In actual fact the basis for selection was not on their psychometrically measured IQ - the children identified were randomly selected. But those designated as ‘gifted’ showed a greater increase in psychometrically measured IQ (approximately 8 IQ points) over the course of a year than the control students, highlighting that teacher and parental expectations can exert a powerful influence on intellectual functioning.

Now, imagine a similar hypothetical experiment where half the participants are assigned to a social category that identified them as mathematically gifted but poor in reading and language proficiency. Even if there were no initial differences in initial starting ability, over the course of their schooling gender stereotypes might well become self-fulfilling prophecies, shaping how children see themselves in relation to intellectual domains like reading and STEM.

2.3.3 Macro-level Cultural Contributions

Most psychobiosocial models of sex differences consider the role of biological and social processes within a given sample (typically but not always drawn from the USA), but neglect to consider the contribution of larger macro-level cultural factors such as cultural beliefs and practices or different educational systems. Miller and Halpern (2013) note that studies examining extremely large datasets (e.g., $n > 100,000$)
have found considerable cross-cultural variability for sex differences in cognitive ability.

Such datasets also provide an excellent opportunity to test theoretical arguments about the origins of sex differences in cognitive ability. Where the source of sex differences are largely the product of internal biological factors (such as those reviewed earlier), the pattern of sex differences should be universal (i.e., all countries show superior female performance or all countries show superior male performance) and largely homogenous. This is the case for sex differences in reading (Guiso et al., 2008; Lynn & Mikk, 2009), and for visual-spatial ability (I. Silverman, Choi, & Peters, 2007). Where the source of sex differences are largely the product of external psychosocial factors, and to the extent that these factors may vary in strength from one country to another, we should see a pattern of sex differences that are more heterogeneous. The heterogeneity may come in two forms: firstly in magnitude (e.g., some countries show negligible or very small effect sizes while others show larger effect sizes), or secondly in direction (i.e., some countries show superior male performance, but others show superior female performance for the same task). By way of example, cross-cultural patterns of sex differences in mathematics and science in adolescence demonstrate both properties. Reilly, Neumann and Andrews (2017) reported data from the TIMSS 2011 wave that are highly heterogeneous in both direction (greater female performance is found in some countries, greater male performance in others) and magnitude (some nations have quite large sex differences in achievement while other nations show negligible gender gaps). This would suggest that mathematics and science outcomes are highly malleable and that under the right environmental and cultural conditions, any outcome is possible.
The two most prominent theoretical perspectives on cultural contributions to the development of sex differences are Eagly and Wood’s (2011) social role theory, and the gender stratification hypothesis (D. P. Baker & Jones, 1993; Riegle-Crumb, 2005). While biological and psychosocial forces may represent proximal factors for individual differences in cognitive ability, these macro-level cultural factors represent a more distal influence in the emergence of group differences between males and females (i.e., population effects, rather than for specific individuals).

**2.3.3.1 Social Role Theory.**

Eagly (1987, 1997) proposed social role theory as a contrast to arguments made by evolutionary psychology. Rather than sex differences in behaviour and cognition being primarily biologically driven by the abilities and limitations of one gender or another, the theory posits that sex differences are largely socially constructed and arise from the historical gendered segregation of labour and responsibilities in society. Eagly and Wood (2016) argue that sex differences in behaviour and cognition reflect sex-role beliefs that, in turn, represent cultural perceptions of women’s and men’s social roles in society. Social role theory proposes two direct mechanisms by which these are realized: firstly, societal divisions between the roles and responsibilities of males and females lead to the formation and perpetuation of gender stereotypes and sex-role beliefs; secondly, sex-roles beliefs act as a self-regulatory process that constrains thought and behaviour in a gender-consistent manner (see Section 2.3.2.3), as well as providing an evaluative framework of expectations for the behaviour of others.

Eagly and Wood (2011) note that in post-industrial societies, men are more likely than women to be employed full time, that they occupy higher status roles and positions of authority, and that women are more likely than men to fulfil caretaking roles either in the home or in the workforce. Observation of these divisions of labour
between the sexes produces expectancies about their underlying disposition and capabilities: because men and women generally assume different social and occupational roles, cultural beliefs develop that men and women possess different traits and capabilities (e.g., men occupy positions of leadership, therefore men are seen as dominant and assertive, whereas women have greater representation in caring professions, therefore they are more caring and nurturing). These expectancies form the basis of consensually-shared beliefs held by society, or gender stereotypes. The gendered division of roles may also exert an influence through observation and modelling (see Section 2.3.2.4), helping to shape a child’s occupational and intellectual aspirations, particularly in male-dominated fields such as STEM or in female-dominated fields such as education, childcare and nursing. Eagly and Wood (2016, p. 459) note that because these sex-roles are seen to “reflect innate attributes of the sexes, they appear natural and inevitable”. To the extent that women’s and men’s roles remain unchanged they will be transmitted generationally, but unlike evolutionary psychological perspectives, social role theory suggests they may be subject to intervention (for example, by increasing the availability and acceptance of counter-examples such as female engineers or male teachers).

All cultures share beliefs about essential differences between males and females (Best, 1982; Williams & Best, 1990), but the magnitude of gender stereotypes does vary across countries. The underlying premise behind social role theory (psychological sex differences are the result of observation of men and women’s social roles) provides a testable hypothesis that can be examined cross-culturally. It predicts that there will be an observable relationship between the representation of women in the workforce and the magnitude of sex differences in cognitive ability. There is some empirical support for such claims. For example, Else-Quest, Hyde and Linn (2010) investigated sex
differences in mathematics achievement conducted for the Programme for International Student Assessment (PISA) 2003 wave. The researchers found that women’s share of higher labour market positions and the percentage of scientific researchers who were female both significantly predicted the magnitude of sex differences in mathematics. It also suggests that directly challenging gender stereotypes (e.g., scientists are male) and increasing availability of counter-examples might also affect positive change in the magnitude of sex differences (Nosek et al., 2009), and is consistent with research showing the importance of female role models and mentoring for increasing representation of women in STEM-related fields (Carli, Alawa, Lee, Zhao, & Kim, 2016).

2.3.3.2 Gender stratification hypothesis.

In a similar manner to social role theory, the gender stratification hypothesis argues that a contributing factor to sex differences in cognitive abilities is gender inequality throughout all levels of society (including occupational roles, educational attainment and political representation). In some research studies, it has been termed the gender segregation hypothesis, or the gender equality hypothesis. The distinction between social role theory and the gender stratification hypothesis is that social role theory makes a causal attribution and provides a mechanism. That is, beliefs about gender stereotypes are derived from observation of the division in society of men’s and women’s roles, and adherence to these gendered social roles and power structures results in sex differences. By contrast, the gender segregation hypothesis is more tentative, and less prescribed. Individual factors such as female representation in parliament or the workforce that show a significant correlation with sex differences are not necessarily causal, but rather are a general proxy for attitudes towards women and gender equality. Women in countries where gender equality is low face additional
barriers to access education or the labour market (e.g., discouragement by parents and teachers to pursue higher education, lack of availability of maternity leave, absence of anti-discrimination protection in hiring practices, etc.) independently of their own ability level and self-efficacy. These are aggregated to produce an objective measure of a nation’s level of gender equality (Else-Quest & Grabe, 2012). Fortin (2005) found that traditional sex-role attitudes of a country are associated with a reduction in female educational participation and their participation in the labour market. Women also encounter fewer female role models in positions of power (especially in highly male-dominated professions), which are particularly important in challenging negative gender stereotypes in fields such as STEM.

Baker and Jones (1993) were the first to investigate gender segregation of roles by examining cross-cultural patterns of sex differences in the Second International Mathematics Study (SIMS), a large international assessment of mathematics achievement that was conducted in 1964. Baker and Jones found medium-sized negative correlations with a range of societal measures of gender segregation in education and the workforce, such that sex differences were smaller with greater representation of women. Such findings should be interpreted cautiously, however, given the small number of countries participating (n = 19), and age of the findings (considerable change in women’s roles may have taken place; sex differences in mathematics may be smaller than for previous generations).

More recently, Riegle-Crumb (2005) repeated the analysis using data from the Third International Mathematics and Science Survey (TIMSS) conducted in 1995. Although there was no association between gender stratification and achievement in maths and science, there were significant associations for attitudes towards mathematics and science (i.e., where there was less segregation, girls’ attitudes towards science and
Another study by Guiso, Monte, Sapienza, and Zingales (2008) conducted a similar analysis but with older children from the Programme for International Student Assessment (PISA), finding that sex differences in mathematics achievement disappear in countries with greater gender equality. However, some researchers have reported a failure to replicate these findings with later waves of PISA data (Tao & Michalopoulos, 2017), leading to some uncertainty over support for the gender segregation hypothesis.

To date, the gender stratification hypothesis has been applied almost exclusively to mathematics and science achievement. Guiso et al. (2008) also examined whether there were similar patterns for sex differences in reading (where females outperform males). They found a positive correlation with gender equality, such that more gender equal nations had larger sex differences in reading.

### 2.3.4 Nash’s Sex-Role Mediation Theory

As outlined above, there are a large number of biological and psychosocial explanations for the development of sex differences between males and females as a group. However as Halpern et al., (2007) have noted, there is also greater within-gender variability than between-gender differences, and substantial overlap between the sexes. Why do some males perform poorly on visual-spatial and quantitative reasoning tasks compared to their male peers, and similarly females with verbal and language abilities? An integrated theory that could explain group differences and individual differences in cognitive ability would be a significant advancement on current explanations, as would a theory that bridged the divide between biological and psychosocial perspectives.

The present course of research examines a promising theoretical explanation for sex differences that integrates aspects of biology, psychosocial and cultural contributions - Nash’s (1979) sex-role mediation theory of cognitive sex differences.
Specifically, Nash (1979) proposed that masculine sex-role identification facilitated the cultivation of visual-spatial reasoning, while feminine sex-role identification encouraged the development of verbal and language abilities (see Figure 2.1). In a review of sex-role identification and cognitive ability, Nash wrote: “For some people, cultural myths are translated into personality beliefs which can affect cognitive functioning in sex-typed intellectual domains” (p. 263). The sex-role mediation theory posited two distinct pathways by which sex-roles might influence the development of cognitive abilities. It was the synthesis of two ideas (performance on a cognitive task could be influenced by the perceived sex-typing of the task, and that sex-role identification could provide additional opportunities to hone and practise one’s talents) offered by Nash (1979) that placed development of cognitive ability in a social context, where sex-role identification encourages or discourages development of intellectual potential. It also acknowledges the interaction between biological and psychosocial factors, in that the sex-role identification process may also be influenced by hormonal expression and prenatal experiences (Knafo, Iervolino, & Plomin, 2005), or by early socialization experiences and cultural stereotype (Chaplin et al., 2005; Eccles et al., 1990; Fagot & Hagan, 1991).

Figure 2.1 – Dual-pathway mechanism of Nash’s sex-role mediation theory
Nash’s (1979) sex-role mediation theory arose from two earlier studies by Milton (1959) and Sherman (1967) into sex-roles and intellectual performance. Milton had hypothesised that sex differences in cognitive performance might be at least partially influenced by the perceived sex-role appropriateness of tasks, and demonstrated that experimentally altering perceptions of problem-solving tasks to be stereotypically masculine or feminine influenced cognitive performance (Milton, 1958). Sherman proposed an entirely different mechanism by which sex-role identification might affect performance. Rather than addressing sex-typing of tasks and sex-role conformity pressures, Sherman (1967) hypothesised that the sex difference between males and females on visual-spatial tasks might simply be the result of differential levels of practice between boys and girls. Many childrens’ leisure pursuits and activities were highly sex-typed at the time (and some remain so today). Activities that promote spatial learning such as mechanical drawing, carpentry, model building, construction blocks, and organised sports provided additional learning experiences that promote the development of visual-spatial reasoning. Boys generally choose to participate in such activities when made available, and typically spend more hours on these activities than girls (Casey, 1996). Sherman used the analogy of a bent twig, in reference to the old adage “as the twig is bent, so shall the tree grow” n.d. Even initially quite tiny biological sex differences in visual-spatial reasoning might interact with environmental experiences that promote learning (the twig), and affecting the direction of intellectual growth over a prolonged period. Subsequently, this has been referred to as the Bent Twig Theory of sex differences (Casey, 1996; Doyle et al., 2012). Robust associations

---

3 Historically, organized sport at the time in the United States was seen as more gender appropriate for boys and this bias was reflected in research conducted at the time (B. Stevenson, 2010). Following legislative change mandating equality of educational opportunities for girls (Title IX legislation), attitudes towards encouragement and support for girls and organized sport improved. Subsequently a moderately strong association between sport and spatial reasoning has been found for males and females (Voyer & Jansen, 2017), supporting part of Sherman’s argument.
have been found between retrospective self-reports on childhood spatial activities questionnaires and visual-spatial performance (Baenninger & Newcombe, 1995; Doyle et al., 2012; Signorella et al., 1989), suggesting that they do provide additional practice and learning opportunities.

Similar arguments had been put forward by other researchers for the development of verbal and language abilities. Kagan (1964) studied the sex-role classification of subjects in school and observed that most students classified reading as being sex-typed as feminine. This view has been replicated in more modern samples (Lynch, 2002; Martino, 2001; Millard, 1997; Shapiro, 1990), with consequential effects for boys and girls on reading motivation and self-efficacy. Like visual-spatial ability, the primary mechanism by which sex-role identification would affect reading skill would be self-selection of activities that lead to differential levels of practice. For example, feminine sex-role identification is associated with more favourable attitudes towards reading, as well as significantly greater amounts of leisure reading (Turner, 1983), giving additional training time on reading and language development. But performance on a verbal or language task might also be affected by the test-taker’s perceptions of the sex-role appropriateness of the task and cultural stereotypes, especially on standardised tests of achievement and grades.

In a meta-analysis of the association of sex-role identification and cognitive performance, Signorella and Jamison (1986) note that only a handful of studies have investigated the relationship between verbal ability and femininity. Most studies have examined reading in children, though these are subject to methodological issues such as small sample sizes and restricted age ranges. Schickendanz (1973) examined a male-only sample of third-grade students drawn from several elementary schools, finding a weak positive association between feminine sex-role identification and reading ability, $r$
Similarly, Dwyer (1974) examined a cross-section of 384 students from Grades 2-12, finding that sex-role beliefs about reading significantly predicted reading ability, even after controlling for sex and grade level. However Nash (1974) found an interaction between sex and sex-role identification in a sample of 207 children (ages 11-14). For girls, feminine sex-role identification was associated with higher reading scores, but surprisingly a significant effect was not found in boys for this sample. Other studies have found similar findings with children and reading motivation (Turner, 1983). More recently McGeown, Goodwin, Henderson and Wright (2011) examined reading motivation and ability in a sample of 182 primary school children from the UK aged 8-11. Students completed the Children’s Sex Role Inventory (CSRI), a questionnaire on reading motivation and interest, as well as completing a standardised test of reading comprehension and survey of attitudes towards reading. They found that sex-role identification was a better predictor of reading performance than biological gender, with feminine identification showing a moderately strong positive association with reading motivation and self-efficacy beliefs. However the study failed to find a meaningful association with reading skill on a standardised test. Finally, a recent study by Ehrtmann and Wolter (in press) examined the effect of gender-role orientation on reading achievement in a nationally representative German sample of secondary students. They found that students who endorsed traditional gender-roles showed lower test performance for reading than those endorsing an egalitarian gender-role orientation. However a limitation of their study was the reliability and validity of their gender-role orientation measure which was operationalized differently to established measures of sex-role orientation.

Taken together, there is tentative evidence of a sex-role mediation effect for reading attitudes, motivation and self-efficacy, but the hypothesis has not been strongly
tested for actual reading ability. Whether there would be transfer effects to performance on other types of language tasks is also unclear, because so few studies have tested the sex-role mediation hypothesis for broader language tasks (Signorella & Jamison, 1986). Only one identified study has employed reliable measures of sex-role identification and verbal ability in a modern sample. Ritter (2004) examined a sample of 79 college students, finding that feminine and androgynous adults scored higher on a verbal fluency measure in males ($d = 1.42$), and females ($d = .99$). Further research is required to determine whether this is the result of differential levels of practice, sex-role appropriateness, or some combination of both.

2.4 Summary of Literature Review Findings

There are four main aspects of cognitive ability that show noteworthy sex differences (see Table 2.1). These include the three domains identified by Maccoby and Jacklin (1974), which were verbal and language abilities, visual-spatial ability, and quantitative reasoning. But later empirical research and scholarship identified that sex differences were also present in memory (recognition and recall) across multiple modalities, though such findings are typically overlooked by broad reviews (e.g., Hyde, 2005). Researchers generally concur that there are robust sex differences for verbal and language abilities as well as for visual-spatial, but many aspects of verbal ability are under-investigated and lack replication with modern samples. However cross-cultural studies show that sex differences in quantitative reasoning are not found in all samples, and may be at least partially influenced by socio-cultural factors.

There is also a wide body of research that investigates the way in which people think and feel about intelligence (in oneself, and in others). Males typically rate their intelligence as higher than do females (male-hubris, female humility effect), but on average both sexes estimate the intelligence of male relatives as being higher than
female relatives. This observation runs contrary to established evidence showing that males and females do not differ in measured psychometric and general intelligence. However, research does show that when asked about specific cognitive abilities, lay persons are fairly accurate in their estimation of actual sex differences in the general population.

A variety of explanations have been offered for the emergence of sex differences in cognitive ability, which have been termed origin theories. These include biological explanations such as the contribution of sex hormones on the brain of the developing foetus, as well as activational effects that are strongest after puberty (a time where the gender gap in sex-typed cognitive abilities widens). Evolutionary psychologists have also argued that there may be genetic differences between biologically female (XX chromosome) and male (XY chromosome) that have been shaped by the division of male and female roles throughout human prehistory. In recent decades though there has been a shift away from biological determinism and towards acknowledgement of the differences in early socialisation experiences of boys and girls, and the ongoing contribution of sex-roles and gender stereotypes. There is also a growing recognition that macro-level cultural factors such as gender equality and implicit gender beliefs have on the intellectual interests and performance of boys and girls. A growing body of literature has identified a variety of mechanisms by which sex differences in specific cognitive abilities are made manifest, and most sex difference researchers endorse a broad psychobiosocial model rather than any single origin theory. One theory that encompasses biological and psychosocial contributions is Nash’s sex role mediation theory of sex differences, which holds that the process of sex-role identification leads to self-selection of activities and interests resulting in differential levels of training in specific cognitive abilities. Additionally, the perceived sex-typing of a given task also
contributes to cognitive performance: when perceived sex-typing of the task is incompatible with an individual’s sex-role identification this may result in lowered performance. Though there is fair support for the sex-role mediation theory on visual-spatial tasks much of the literature is dated and few studies have investigated verbal and language abilities.

Origin theories of sex differences are important as they may identify targets for educational and psychosocial interventions, but are difficult to test empirically – when a study or meta-analysis reports a difference between males and females, the relative contribution of biological and psychosocial factors cannot be determined, especially if a sample is selective and not representative. But the tenability of such origin theories can be tested cross-culturally by examining variability in the direction and magnitude of the gender gap. A strong biological contribution would be consistent with observed differences in verbal and language abilities, as well as visual-spatial reasoning. However quantitative reasoning as measured by educational achievement in mathematics and science is highly culturally variable: in some nations females score higher than males, while in others there may be no difference whatsoever or that males score higher than females. This pattern of observations would be more consistent with psychosocial origin theories. At present, there exists no cross-cultural studies of sex differences in memory.

References


Turner, M. A. (1983). *Sex-role Identity Of Children And Its Relationship To The Sex-role Identity Of Parents And Teachers, And To Reading Interests, Attitudes Toward Reading, And Level Of Reading Proficiency.* (Doctor of Education thesis), Lehigh University, Bethlehem, Pennsylvania. Retrieved from ProQuest Dissertations and Theses database. (Proquest ID 303270815)


Chapter 3 – Gender Differences in Spatial Ability

This chapter provides a literature review on sex differences in visual-spatial ability, and contains supplementary material to the general literature review in Chapter 2.

This work has been published as:


Permission for inclusion of the final paper has been granted by the publisher, Springer Nature. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
Conducting literature review
Writing chapter

(Signed) ________________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ________________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann

(Countersigned) ________________________________ (Date) : 1/12/18
Associate Supervisor Glenda Andrews
Chapter 4 – Sex Differences in Mathematics and Science Achievement

This chapter reports on a meta-analysis of archival data from a large nationally representative assessment of student achievement in the domains of mathematics and science conducted in the United States. Student achievement in Grades 4, 8, and 12 have been periodically assessed across a sufficiently long time period that it is possible to test for temporal trends such as the predicted decline in gender gaps with changes in the relative status of men and women in society.

This chapter includes a co-authored paper that has been published as:

doi: 10.1037/edu0000012

Permission for inclusion of the final paper has been granted by the publisher, American Psychological Association. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
- Data collection from archival sources
- Statistical Analysis
- Writing chapter

(Signed) ____________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ____________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann

(Countersigned) ____________________________ (Date) : 1/12/18
Associate Supervisor Glenda Andrews
Chapter 5 – Sex Differences in Reading and Writing

This chapter reports on a meta-analysis of archival data from a large nationally representative assessment of student achievement in reading and writing conducted in the United States as part of the National Assessment of Educational Progress. Student achievement in Grades 4, 8, and 12 have been periodically assessed across a sufficiently long time period that it is possible to test for temporal trends such as the predicted decline in gender gaps with changes in the relative status of men and women in society. Large and pervasive sex differences were found for writing achievement, especially gender ratios at the tails. Some somewhat smaller sex differences were found for reading, and there are twice as many boys as girls failing to attain basic literacy at the lower left-tail of the ability distribution.

This chapter includes a co-authored paper that has been published as:


Permission for inclusion of the final paper has been granted by the publisher, American Psychological Association. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
- Data collection from archival sources
- Statistical Analysis
- Writing chapter

(Signed) ____________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ____________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann

Copyright
(Countersigned) ____________________________ (Date) : 1/12/18
Associate Supervisor Glenda Andrews
Chapter 6 – Cross-Cultural Patterns of Reading, Mathematics and Science Literacy

This chapter reports on a meta-analysis of student testing data from the 2009 wave of the Programme for International Student Assessment (PISA), a large-scale educational assessment of student’s reading, mathematics and science literacy across all OECD members and a number of partner nations. This study reports data from 65 nations. Consistently across all nations, girls outperform boys in reading literacy, $d = -0.44$. Boys outperform girls in mathematics in the USA, $d = +0.22$ and across OECD nations, $d = +0.13$. For science literacy, while the USA showed the largest gender difference across all OECD nations, $d = +0.14$, gender differences across OECD nations were non-significant, and a small female advantage was found for non-OECD nations, $d = -0.09$. Across all three domains, these differences were more pronounced at both tails of the distribution for low- and high-achievers. Considerable cross-cultural variability was also observed, and national gender differences were correlated with gender equity measures, economic prosperity, and Hofstede’s cultural dimension of power distance. Educational and societal implications of such gender gaps are addressed, as well as the mechanisms by which gender differences in cognitive abilities are culturally mediated.

It has been published as


Copyright statement:

It was published in accordance with the Creative Commons Attribution (CC_BY) license, and copyright was retained by the author.
Chapter 7 – Meta-Analysis of Sex-Role Mediation Effect for Visual-Spatial Ability

This study reports a meta-analysis of the sex-role mediation effect for visual-spatial ability. This chapter includes a co-authored paper that has been published as:


Permission for inclusion of the final paper has been granted by the publisher, Springer. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
- Data collection from archival sources
- Statistical Analysis
- Writing chapter

(Signed) ________________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ________________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann
Meta-Analysis Summary and Prelude to Empirical Studies

The previous set of chapters sought to examine whether the previously observed sex differences in specific cognitive abilities (verbal and quantitative reasoning) would still exist in contemporary samples, or alternatively, if they’d been eliminated as claimed by Feingold (1988), Hyde (2005), as well as Caplan and Caplan (1997, 2016). Additionally, it sought to contextualise that difference by evaluating the magnitude of observed sex differences, and determine if they were large enough to have practical importance. For quantitative reasoning, small but not trivial mean sex differences were found for mathematics and more substantial gender gaps in high achievers. Somewhat larger mean sex differences were found for science achievement, and again a sharp disparity in the tail ratios for high achievers. For verbal and language abilities, substantial sex differences were found for reading and writing with a developmental trend observed with age/years of schooling. Examination of tail ratios also showed substantial gender gaps in low- and high- achievers.

Collectively these studies provided a rationale for further investigation to evaluate support for the sex-role mediation hypothesis. Signorella and Jamison (1986) had conducted a meta-analysis on the association between masculinity and visual-spatial ability, but the literature was now dated. For this reason, we produced a meta-analysis of the association between masculine sex-role identification and visual-spatial ability with more recently collected data. Additionally, this was useful in contextualising the expected effect size for statistical power calculations in the empirical study that follows.
Chapter 8 – Empirical Study 1 – Sex and Sex-Role Differences in Specific Cognitive Abilities

“If women are expected to do the same work as men, we must teach them the same things.” – Plato, The Republic.

This study reports the empirical study into sex and sex-role differences in verbal and visual-spatial abilities. Specifically it tests Nash’s (1979) sex-role mediation hypothesis in a modern sample, finding support for both predicted tranches (verbal and visual-spatial ability) across a range of tasks. This has been published as:


Permission for inclusion of the final paper has been granted by the publisher, Elsevier. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
- Data collection from archival sources
- Statistical Analysis
- Writing chapter

(Signed) ________________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ________________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann

(Countersigned) ________________________________ (Date) : 1/12/18
Associate Supervisor Glenda Andrews
Chapter 9 – Empirical Study 2 – Effect of Task-Labelling, Stereotype Threat, and Sex-Role Identification on Cognitive Performance

“We see the world, not as it is, but as we are—or, as we are conditioned to see it.”

Stephen R. Covey, The 7 Habits of Highly Effective People

This study reports an empirical study investigating the effect of task labelling, stereotype threat induction, and sex-role identification on cognitive performance in visual-spatial and verbal ability tasks in a sample of 150 women. It also concurrently measures sex-role identification in participants, in order to replicate the findings of Chapter 8.
“What does my performance on this test say about me?” : Effect of Task-Labelling, Stereotype Threat, and Sex-Role Identification on Cognitive Performance

Abstract

Sex differences in cognitive ability are explained by the sex-role mediation hypothesis as arising from the development of sex-typed personality traits and behaviors. Other researchers claim sex differences in latent ability do not exist, but instead reflect diminished performance in the face of stereotype threat or gender conformity pressures. A sample of 150 women was recruited to investigate the effect of task-labelling, stereotype threat and sex-role identification on cognitive performance. Initially the women were randomly assigned to either a masculine or feminine task-labelling condition before completing a spatial visualization task. Next, the women were randomly assigned to either a stereotype threat or control condition, and then completed a mental rotation and verbal fluency task. Results on visual-spatial tasks showed effects of task-labelling and stereotype threat, as well as sex-role differences consistent with the sex-role mediation hypothesis. Additionally, sex-role differences were found for a verbal fluency task but there was no effect of stereotype lift. The results suggest that the sex-role mediation effect observed in previous studies for visual-spatial tasks reflects an enduring trait, but can be moderated by task-labelling and salience of gender stereotypes.

Keywords: sex differences, stereotype threat, sex-role mediation hypothesis, spatial ability, verbal ability, gender priming
“What does my performance on this test say about me?”: Effect of Task-Labelling, Stereotype Threat, and Sex-Role Identification on Cognitive Performance

The topic of sex differences in cognitive abilities has commanded the interest of psychologists and researchers since the beginning of our field. It has also captured the curiosity of parents, educators, and the media, due to the important educational and social implications (Eagly, 1996; Halpern, 1997). The social significance of sex differences include the underrepresentation of women in science and technology fields (Carli, Alawa, Lee, Zhao, & Kim, 2016; Halpern et al., 2007), disparities between boys and girls on standardized tests of reading and writing (Hedges & Nowell, 1995; Reilly, 2012; Reilly, Neumann & Andrews, 2018), as well as putative claims that men and women are inherently “different” and would benefit from single-sex education environments (for a critical review of evidence see Halpern et al., 2011).

While sex differences in most types of cognitive ability are relatively small in magnitude (Hyde, 2005), two exceptions to this general rule are verbal and visual-spatial abilities (Maccoby & Jacklin, 1974). On average females score higher on tasks involving language and verbal ability while males tend to score higher on tasks of visual-spatial ability (Halpern, 2011; Voyer, Voyer, & Bryden, 1995). However, there is vigorous debate amongst researchers over the extent to which the observed differences reflect biological and psychosocial factors, necessitating the need for further research. Additionally, it is unclear whether observed differences reflect actual sex differences in latent ability or rather instead are an artefact of the testing environment and situational factors (Massa, Mayer, & Bohon, 2005), such as stereotype threat (Flore & Wicherts, 2015; Steele, 1998).

One line of enquiry into the causes of sex differences highlights the role of individual differences in sex-role identification on the performance of gender-typed
cognitive tasks. The sex-role mediation hypothesis proposes that sex differences between males and females on cognitive tests stem from the development of stereotypically masculine and feminine personality traits and behaviours (Nash, 1979; Reilly, Neumann, & Andrews, 2016). Although the early socialization experiences of boys and girls typically differ (Leaper & Friedman, 2007), there is also considerable individual variation in the degree to which children acquire stereotypically masculine and feminine personality traits, beliefs, and behaviours – a process referred to as sex-typing (Kohlberg, 1966; Martin & Ruble, 2004). Highly sex-typed persons are motivated to keep their behaviour and self-concept consistent with traditional gender norms (Bem & Lenney, 1976), including performance in academic domains (Carli et al., 2016; Nosek, Banaji, & Greenwald, 2002; Steffens & Jelenec, 2011). Others may integrate aspects of both masculine and feminine identification into their personality, termed psychological androgyny (Bem, 1984; Spence & Buckner, 2000).

Many persons acquire stereotypically masculine or feminine sex-roles consistent with their biological gender, resulting in group differences between males and females being observed. However the sex-mediation hypothesis also offers an explanation for individual differences in performance, which is important because within-gender variability on cognitive tasks is often greater than between-gender variability (Halpern et al., 2007). Several studies have investigated sex-role identity and cognitive performance (e.g. Saucier, McCreary, & Saxberg, 2002). Few studies have examined interactions between sex-role identification and situational factors of the testing environment such as the perceived sex-role appropriateness of a task, or activation of stereotype threat. The aim of the present study was to investigate the effects of sex-role identification, and its relationship to these two intrapersonal factors on women’s cognitive performance.
Overview of Gender Differences in Specific Cognitive Abilities

While it is generally regarded that men and women do not differ in general intelligence (Jensen, 1998; Neisser et al., 1996), sex differences do exist for more specific cognitive abilities. The largest sex differences are found in visual-spatial ability where males as a group perform better than females (Voyer et al., 1995). But there are also appreciable differences in verbal fluency (Gauthier, Duyme, Zanca, & Capron, 2009; Weiss, Kemmler, Deisenhammer, Fleischhacker, & Delazer, 2003) and language tasks such as reading and writing (Berninger, Nielsen, Abbott, Wijsman, & Raskind, 2008; Lynn & Mik, 2009; Reilly, 2012), where females score significantly higher than males on average.

These two domains are important to cultivate because visual-spatial ability predicts later development of mathematical and scientific aptitude (Duffy, Sorby, Nozaki, & Bowe, 2016; Kersh, Casey, & Young, 2008; Verdine, Golinkoff, Hirsh-Pasek, & Newcombe, 2017; Wai, Lubinski, & Benbow, 2009), while verbal and language proficiency are crucial skills for occupational success and higher socioeconomic status (Kutner et al., 2007; Ritchie & Bates, 2013). Females on average score slightly lower in standardized tests of mathematics and science (Ganley & Lubienski, 2016; Reilly, Neumann, & Andrews, 2015), with more pronounced gender gaps in the proportion of high achievers at the upper-right-tail of the ability distribution (Wai, Cacchio, Putallaz, & Makel, 2010). Males also score lower than females on standardized tests of reading and writing (Lietz, 2006; Willingham & Cole, 1997). But there is also considerable variability in performance within both males and females as groups. For example, some men score poorly on tests of mathematics and science, while some women score considerably higher than their male peers on such tests. Halpern et al. (2007) observed that within-gender variability is, in fact, greater than between-gender
variability. Any theoretical explanation for the origin of sex differences in cognitive ability should thus consider the role that individual differences play in the emergence of group differences between males and females.

**Sex-Role Mediation as an Explanation for Sex differences**

Nash’s (1979) sex-role mediation hypothesis proposes that within-gender and between-gender variability arises as a result of differences in sex-role identification when performing stereotypically masculine or feminine cognitive tasks. Specifically, Nash (1979) theorized that masculine identification leads to the cultivation of visual-spatial ability, while feminine identification promotes the acquisition of verbal and language abilities (see Figure 9.1). Sex-role identification occurs early during childhood and represents an enduring personality trait that is relatively stable over time (Hyde, Krajnik, & Skuldt-Niederberger, 1991; Martin & Ruble, 2004). While there have been significant changes in the relative roles of men and women (Donnelly et al., 2015), recent research shows the stability of gendered stereotypes over time (Haines, Deaux, & Lofaro, 2016).

![Figure 9.1. Sex-role mediation theory of cognitive abilities. Masculine sex-role identification is associated with increased spatial experiences while feminine sex-role identification provides additional opportunities to develop verbal and language proficiency.](image-url)
Nash’s (1979) sex-role mediation hypothesis was based on earlier work by Sherman (1967) and others into how sex-role identification can result in differential learning and practice experiences in early childhood, as well as self-selection of hobbies and interests that promote either visual-spatial or verbal ability. It also reflected research into the self-concept and gender-schema theory, which argued that rigidly sex-typed individuals are highly motivated to keep their actions consistent with their self-concept. Thus they may show less motivation to persevere in the face of difficult and challenging content when performing a sex-typed cognitive task that is incompatible with their self-concept, and may be more susceptible to negative gender stereotypes about ability and overt gender bias (LaCosse, Sekaquaptewa, & Bennett, 2016; Robnett, 2015). In doing so, Nash extended Sherman’s work to place cognitive development in a social context, where appraisal of task characteristics, prior practice and experience with the task, and knowledge of gender stereotypes about ability may all contribute to sex-typed cognitive abilities.

The sex-role mediation hypothesis also predicts individual differences in actual ability, due to self-selection of stereotypically masculine and feminine leisure activities and intellectual interests. For example, engagement in stereotypically masculine leisure activities such as construction block play, video gaming, and model-making in childhood and adolescence is correlated with adult performance on visual-spatial tasks (Baenninger & Newcombe, 1995). Feminine sex-role identification is associated with more positive attitudes to reading (McGeown, Goodwin, Henderson, & Wright, 2011), and to engagement in activities such as reading for leisure, creative writing and journal keeping (Athenstaedt, Mikula, & Bredt, 2009; McHale, Kim, Whiteman, & Crouter, 2004). These activities provide further opportunities to practice and cultivate specific
cognitive skills. Individual differences in sex-role identification may lead to differential levels of practice and experience with sex-typed cognitive tasks.

Several decades have passed since Nash’s (1979) sex-role mediation hypothesis was first proposed. There have been changes in the status and roles of women during this period as well as changes in cultural perceptions of masculinity and femininity (Auster & Ohm, 2000). These changes raise the question of whether earlier research findings will be replicated with modern samples (Eagly & Diekman, 2003). Signorella and Jamison (1986) found a positive association between masculine sex-roles and visual-spatial performance in a meta-analysis of research findings. Subsequently, a meta-analysis by Reilly and Neumann (2013) examined the contribution of masculine sex-roles to mental rotation performance and found the relationship still holds with modern samples. However, few studies had examined the contribution of femininity to verbal ability and language tasks.

More recently an empirical study by Reilly, Neumann, and Andrews (2016) tested the sex-role mediation hypothesis across a range of visual-spatial and verbal language tasks. Masculinity was associated with higher performance for visual-spatial tasks, while femininity predicted performance on verbal and language tasks. The researchers employed tests of statistical mediation, finding that sex-role identification was a mediator between sex and performance on cognitive tasks. While such research documents the sex-role mediation effect in modern samples, it leaves unanswered the question of whether other factors might also contribute to the observed sex differences in performance. Two situational factors previously shown to affect performance on sex-typed cognitive tests are task-labelling and stereotype threat.
Sex-typing of Cognitive Tests and Task-labelling

One explanation for reduced performance may be a perceived conflict between the sex-typing of a cognitive task as inherently masculine or feminine, and the individual’s sex-role identification. Kagan (1964) noted that from an early age, children begin to classify everyday tasks and behaviours as being either masculine or feminine. Reading is stereotypically regarded as feminine while visual-spatial and mathematical tasks are perceived as masculine (Hyde & Lindberg, 2007; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). When attempting to complete a sex-typed task that is incompatible with one’s sex-role identification, the test-taker may be less motivated and may be easily discouraged when encountering difficult material.

Two studies have attempted to manipulate individuals’ appraisals of cognitive tasks through labelling and to examine the effect on performance. Brosnan (1998) had participants complete the Group Embedded Figures Test (GEFT; Witkin, 1971), after providing them with one of two sets of instructions – one that emphasized the visual-spatial nature of the task, and the other that described it as a test of empathy and perspective-taking. The ambiguous nature of the test and novelty of the stimuli gave such descriptions face validity. Although there was no effect of labelling in males, females in the empathy condition performed significantly better than those in the spatial condition. A second study by Massa, Mayer, and Bohon (2005) replicated the labelling effect on GEFT performance in a female college-aged sample, and also found that motivation was higher when the test was portrayed as being feminine in nature.

Gender Stereotypes, and Stereotype Threat

While the sex-role appropriateness of the task may well be a factor in general performance (e.g. Massa et al., 2005), an alternate explanation for reduced performance is that by describing the task as measuring visual-spatial ability, knowledge of gender
Chapter 4 – Sex Differences in Mathematics and Science Achievement

This chapter reports on a meta-analysis of archival data from a large nationally representative assessment of student achievement in the domains of mathematics and science conducted in the United States. Student achievement in Grades 4, 8, and 12 have been periodically assessed across a sufficiently long time period that it is possible to test for temporal trends such as the predicted decline in gender gaps with changes in the relative status of men and women in society.

This chapter includes a co-authored paper that has been published as:


doi: 10.1037/edu0000012

Permission for inclusion of the final paper has been granted by the publisher, American Psychological Association. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
  - Data collection from archival sources
  - Statistical Analysis
  - Writing chapter

(Signed) _________________________________ (Date) : 1/12/18
David Reilly

(Countersigned) _______________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann

(Countersigned) _______________________________ (Date) : 1/12/18
Associate Supervisor Glenda Andrews
10.1 Introduction

10.1.1 Overview of Gender Differences

The existence of gender differences in cognitive ability is a controversial topic. Nevertheless, researchers in psychological and the social sciences widely acknowledge that males and females differ in spatial ability (Halpern and Collaer 2005; Kimura 2000). Indeed, it is one of the most robust and consistently found phenomenon of all cognitive gender differences (Halpern 2011; Voyer et al. 1995). While there is individual variability within each gender, on average males score higher than females on tests that measure visual-spatial ability. However, there is considerable debate over just how large the differences between males and females are. Researchers also differ in their perspectives on the origins of the gender differences, including the relative contributions of biological, social and cultural factors. This chapter provides an overview of the research literature, as well as covering the developmental and educational implications for children.

Many researchers posit that early expertise in spatial ability in children lays down a foundation for the development of quantitative reasoning, a collective term encompassing science and mathematics. These researchers argue that the early differences in spatial ability have important implications for student achievement in STEM (science, technology, engineering and mathematics) subjects, and may partially explain the underrepresentation of women in science. However, while some
children may be naturally gifted in spatial ability, there is a large body of research showing that spatial proficiency can be improved through relatively brief interventions. A growing number of educational psychologists have argued that early education of spatial intelligence is necessary as a matter of equity for all students, and that it may offer substantial benefits for the later development of mathematical and scientific skills across all ability levels (Halpern et al. 2007). We review interventions aimed at increasing spatial aptitude, and the role of parents and teachers in encouraging the development of these abilities.

### 10.1.2 What Is Spatial Ability?

The term “spatial ability” (also referred to in some research as visuospatial or visual-spatial ability) encompasses a range of different skills and operations, so it is important to clearly define the term. Laypeople can sometimes use the term very loosely, covering anything from block building assembly to reading maps and navigating one’s way around the city streets. Such tasks often incorporate additional (non-spatial) processes, including memory and general problem solving skills. Psychologists and cognitive researchers apply the term spatial ability to tasks that are intended to measure specific cognitive processes in isolation. Linn and Petersen (1985, p. 1482) defined spatial ability as the “skill in representing, transforming, generating and recalling symbolic, non-linguistic information”. More generally, it is the ability to perceive and understand spatial relationships, to visualize spatial stimuli such as objects, and to manipulate or transform them in some way – such as mentally rotating an object to imagine what it might look like viewed from a different angle or perspective. Spatial ability is crucial to a wide variety of traditional occupations including architecture, interior decorating, drafting, aviation, as well as a growing number of new and emerging occupations in the science and technology fields.

Spatial ability encompasses a broad range of cognitive processes, with the size of gender differences varying depending on the type of task (Voyer et al. 1995). When measuring spatial ability, some tasks measure global spatial skills such as wayfinding and navigation in virtual environments or outside the laboratory (Lawton and Kallai 2002). More commonly, specially designed tasks are employed to tap one or more spatial components in isolation. Linn and Petersen (1985), in a pioneering review of the literature, outlined three distinct categories of spatial ability. Firstly, we have spatial perception, which involves perceiving spatial relationships. A commonly employed task of spatial perception is Piagetian Water Level Task, which requires individuals to draw the waterline on a variety of containers or bottles that have been tilted a certain number of degrees (see Fig. 10.1). Another is the Judgment of Line Angle and Position test (JLAP), which requires subjects to correctly judge the orientation of a series of tilted lines (see Fig. 10.2).

The second category of spatial tasks is mental rotation. Tasks measuring mental rotation involve requiring individuals to mentally rotate spatial objects to see how they would look from a different angle or perspective (see Fig. 10.3). Mental rotation
tasks usually involve three dimensional stimuli (Kimura 2000), though some tasks use less complex two dimensional stimuli (Prinzel and Freeman 1995).

The third category of spatial ability is spatial visualization which involve more complicated multistep manipulations of spatial information in order to reach a solution. These tasks often incorporate some element of spatial perception and mental rotation. They are distinguished by having multiple solution strategies for reaching a solution. Common tests of spatial visualization include the Embedded Figures

Fig. 10.1 In the Piaget water level task (Vasta and Liben 1996), subjects are presented with a container of liquid (left), with varying quantities of fluid. The container is then tilted adjacent to the horizontal plane. Subjects must then draw a line to indicate the probable water line in each of these containers.

Fig. 10.2 Representative stimuli for judgement of the Judgment of Line Angle and Position test (JLAP; Collaer et al. 2007). Subjects must match the orientation of stimuli lines (left) to a reference array (right). The correct answers from left to right are 2, 4 and 9.

Fig. 10.3 Sample stimuli from the Vandenberg mental rotation task (Vandenberg and Kuse 1978). Subjects must locate both instances of the target shape (left) amongst the four possible choices. Two of the choices are mirror image distractors. To answer the question correctly, both targets must be located. The correct answer is 1 and 3 (From Peters and Battista (2008). Used by permission).
Test (EFT; see Fig. 10.4), which requires individuals to search for a target shape within a more complex picture of geometric shapes and to ignore distracting visual information. Another task is the Paper Folding task, which requires individuals to visualize how a sheet of paper would appear if it were folded in a certain way and then one or more holes were punched through the folded sheet. Individuals must indicate how the unfurled paper would appear and indicate the position of dots from a series of possible answers (see Fig. 10.5).

Some researchers have proposed a fourth category called spatiotemporal ability, which involves making time-to-arrival judgments or tracking the movement of an object through space (Hunt et al. 1988). Such tasks are computer administered in order to accurately measure response times and determine whether there are discrepancies between projected and actual arrival time (see Fig. 10.6). Other tasks involve directing the path of multiple objects concurrently (see Fig. 10.7; Contreras et al. 2001, 2007). However, it is unclear whether the gender difference observed with these tasks is necessarily spatial in nature, because there is some evidence that

![Target shape](image1)

![Stimuli Item](image2)

![Target shape](image3)

![Stimuli Item](image4)

Fig. 10.4 Spatial visualization items representative of those used in embedded figures tasks (Witkin 1971). Subjects are asked to locate a target shape (shown on the left) within a more complex picture (right).
males are more accurate in time perception generally (Hancock and Rausch 2010; Rammsayer and Lustnauer 1989).

10.1.3 Statistical Methods for Evaluating Gender Differences in Research

Experiments in psychology make heavy use of sampling, as it would be impractical to collect a measurement from every member of a given target population.

When a sufficiently large number of people are recruited, statistical tests can be performed to determine the probability that the observed group differences are due to chance, or whether they are likely to be found again if the experiment was repeated. If the probability that the results of the study occurred by chance is very low, the result is said to be statistically significant. Because research involves volunteer participants giving up their valuable time, and the time of the investigator to supervise data collection, researchers generally seek to minimise the number of participants involved. When extremely small sample sizes are recruited for a study.
it may be lacking in statistical power (the ability to detect a statistically significant effect in a given sample, if indeed the effect in question is genuine). Furthermore, samples may differ in important characteristics, such as age, socioeconomic status, level of education, which may affect the study outcomes, serving to increase or diminish the magnitude of any group differences between males and females. By pooling the data from many studies, statistical power is increased and the researcher can arrive at a more reliable estimate of the true size of a given effect than could be reached from any individual study.

Meta-analysis is a statistical technique employed to summarize research findings across studies. Meta-analysis uses statistical methods to quantify effects across studies in an open and transparent manner, rather than simply comparing the tally of positive to negative studies (referred to as ‘vote counting’) or presenting a subjective interpretation of the scientific literature. For example, a selective review of spatial literature by Caplan et al. (1985) made the surprising claim that gender differences in spatial ability were diminishing and were no longer reliably found. A subsequent meta-analysis by Linn and Petersen (1985) provided strong quantitative evidence in a review of the entire published literature of the time that refuted such claims. Statistical techniques and software have advanced sufficiently in recent times so that it is now possible to test additional hypotheses about potential moderators, such as whether gender differences are diminishing in size across decades, or

Fig. 10.7 Dynamic spatial ability requires subjects to steer two concurrently moving objects to a fixed destination point by clicking on the turn left and turn right buttons. Arrows show motion path of the black and white dots. Representative of the Spatial Orientation Dynamic Test – Revised (SODT-R; Contreras et al. 2007)
whether gender differences are present at certain developmental ages (such as childhood and adolescence).

When comparing two groups (such as males and females), the size of the effect in question is represented using a metric. A commonly used metric is Cohen’s $d$, which represents the mean difference between two groups divided by the pooled standard deviation. The use of a common metric facilitates comparisons across different types of tests and samples, in a way that just reporting the mean difference could not. Cohen (1988) offered a set of guidelines for interpreting the magnitude of these group differences, suggesting that an effect size of $d < 0.20$ could be considered a “small” effect, values of approximately 0.50 could be considered medium in size, and values of 0.80 or greater would be considered large in magnitude. These benchmarks offer even the non-statistician assistance in determining whether the effect in question is practically significant, holding research to a higher standard than statistical significance alone.

### 10.1.4 How Large Are Gender Differences in Spatial Ability?

The meta-analytic review conducted by Voyer et al. (1995) represented the most comprehensive meta-analysis of the research on gender differences in spatial ability published at that time. The review categorised tasks by age, comparing children (under 13 years), adolescents (13–18 years), and adults (over 18 years). Mental rotation tasks showed the largest gender differences ($d = 0.33$ for children, $d = 0.45$ for adolescents and $d = 0.66$ in adults) followed by spatial perception ($d = 0.33$ for children, $d = 0.43$ for adolescents and $d = 0.48$ in adults). Spatial visualization showed the smallest gender differences ($d = 0.02$ in children, growing to 18 for adolescents and $d = 0.23$ in adults). By Cohen’s guidelines, these would be medium-sized gender differences for mental rotation and spatial perception and in the case of spatial visualization tasks, relatively small. Contrary to earlier claims (e.g. Caplan et al. 1985), there is little substantive evidence that gender differences in visual spatial ability have greatly diminished over time though. Furthermore the gender differences follow a developmental progression from relatively small gender differences in childhood towards much larger gender differences in adolescence and adulthood. Though a meta-analysis has not yet been conducted on the type of spatial task called spatiotemporal ability, effect sizes in such studies typically fall in the medium to large range also (Halpern 2000).
10.1.5 When Are Gender Differences in Spatial Ability First Observed?

Gender differences in spatial ability are observed early. Children in primary school show meaningful differences across a range of spatial tasks including mental rotation and spatial transformation (Lachance and Mazzocco 2006; Levine et al. 1999). Indeed, some studies have even observed small sex differences in young infants when simplified tests of spatial reasoning are employed (Moore and Johnson 2008; Quinn and Liben 2008). However, the gender gap in spatial ability does appear to widen around the time of puberty, which some had claimed supported arguments for a biological and hormonal contribution. Correlation by itself does not necessarily prove causation though, as there may be other factors that co-vary with puberty. For example, as developmental researchers would also point out, this is time of increased gender conformity and strengthening of sex-roles (Ruble et al. 2006), as well as greater gender differentiation in play and leisure activities which provide opportunities to practise spatial skill (Baenninger and Newcombe 1989). Even after puberty the gender gap continues to widen, with somewhat larger effect sizes found in adults than adolescents. There is evidence that input and practice is required to fully develop spatial ability (Baenninger and Newcombe 1995), and the increase noted in puberty and in later adulthood may reflect the accumulation of social influences across time rather than the influence of hormonal changes.

10.2 Spatial Ability and Quantitative Reasoning

Spatial ability is thought to underpin the development of quantitative reasoning skills such as mathematics and science (Nuttall et al. 2005; Uttal et al. 2013b), which are important educational objectives. Factor analysis (a statistical technique used to investigate the relationship between tests) of cognitive ability tests show high loading for mathematical performance against a spatial factor (Bornstein 2011; Carrol 1993; Halpern 2000). Wai et al. (2009) note that a large body of research over the course of over 50 years has established that spatial ability plays a crucial role in stimulating the development of quantitative reasoning skills. For example, spatial reasoning is important for understanding diagrams of complex scientific concepts and principals, but individual differences in spatial ability predict learning outcomes with such media in physics and chemistry (Höffler 2010; Kozhevnikov et al. 2007; Wu and Shah 2004). When engaging in complex problem-solving tasks in science and mathematics, students who use spatial imagery and diagrams perform better than students using verbal strategies (Spelke 2005), and growth in spatial working memory is positively correlated with mathematics proficiency (Li and Geary 2013).

Furthermore, performance on measures of spatial ability are predictive of future scholastic achievement in mathematics and science, even many years later (Uttal et al. 2013b). Shea et al. (2001) reported the results of a 20 year longitudinal study
that followed children from seventh grade through to the age of 33. They found that individual differences in spatial ability measured in adolescence predicted educational and vocational outcomes two decades later, even after controlling for pre-existing mathematical and verbal abilities.

Another study by Casey et al. (1995) examined a large sample of U.S. adolescents preparing to sit the Mathematics Scholastic Aptitude Test (SAT-M) for college entry, an important prerequisite for entry into further education in mathematics and science. Performance on the Vandenberg Mental Rotation Task successfully predicted SAT-M entrance scores, even after controlling for general scholastic ability (Casey et al. 1995). Although still significant for males, the relationship between spatial ability and mathematics achievement was stronger for females suggesting that girls may be particularly disadvantaged by deficits in spatial reasoning. Casey et al. suggest that spatial ability acts as an important mediator in the gender gap in STEM achievement. Furthermore, they found that higher spatial ability was associated with greater self-efficacy beliefs about learning mathematics (Casey et al. 1997). Attitudes may exert a powerful influence on whether students decide to undertake further classes in mathematics and science (Ferguson et al. 2015; Simpkins et al. 2006), suggesting that there may be motivational effects as well as cognitive effects when spatial competencies are improved.

### 10.2.1 Importance of Spatial Ability for STEM

Educators, scientists, and policy makers acknowledge the importance of increasing mathematical and science literacy proficiencies for students generally. There is also evidence to suggest that the early gender differences in spatial ability may contribute to the later emergence of gender differences in mathematics and science (Ceci et al. 2009; Wai et al. 2009). Examination of historical scholastic achievement scores in the U.S. by Hedges and Nowell (1995) found that males, on average, have higher achievement scores in mathematics and science. Furthermore, when we examine the extreme right tail of the ability distribution, the gender gap is considerably larger. More recently, studies on data from the federal National Assessment of Educational Progress (NAEP) in the United States replicated these findings. For example, Reilly et al. (2015) observed small but stable mean gender differences in mathematics and science achievement and that at the higher levels of achievement boys outnumber girls by a ratio of 2:1 (Reilly et al. 2015). However gender gaps in maths and science are not inevitable. International assessments of educational achievement find that in some countries, females actually outperform males to a significant degree in mathematics and science (Else-Quest et al. 2010; Guiso et al. 2008; Reilly 2012).

A number of researchers have proposed that in order to address the gender gap in mathematics and science achievement, it is necessary to first address the gender gap in spatial ability (Halpern 2007; Newcombe 2007). Fortunately spatial ability is not a fixed and immutable trait (see the section “Interventions for Training of Spatial Ability”). In a review of educational research on gender difference, Hyde and
Lindberg (2007) argued that even a mild increase in spatial ability might have “multiplier effects in girls’ mathematical and science performance” (Hyde and Lindberg 2007, p. 29). This is an important goal as a matter of gender equity, but we can also see substantial improvements of training for males as well. In a review of the developmental and educational research on spatial ability and STEM and the American educational system, Uttal et al. (2013b) argue that including spatial thinking in the science curriculum could substantially increase the number of students capable of pursuing STEM careers. Given that in many developed countries there are shortages within STEM occupations, addressing spatial proficiency in early education may be an important tool for improving overall mathematics and science literacy.

10.3 Theoretical Perspectives on Origins of Gender Differences

Halpern and Collaer (2005) described gender differences in spatial ability as some of the largest found for any cognitive task, raising the important question as to its developmental origins. Why do males on average outperform females on spatial tasks? Past approaches to this question have emphasized biological factors as well as social factors, cultural influences, and life experiences. It is unlikely that there is one single factor that can adequately explain the magnitude of the gender gap for spatial ability. Most gender difference researchers would acknowledge both biological and social forces contribute to their development, embracing a biopsychosocial model of gender differences (Halpern and Tan 2001; Hyde 2014). While there may be biological factors that predispose an individual to greater or lesser proficiency on spatial tasks, it must be remembered that they are not immutable. Full development of such skills requires practice and experience, and both males and females can make significant gains with training.

10.3.1 Evolutionary and Genetic Factors

Evolutionary psychology seeks to make sense of gender differences in human cognition by considering the role of evolutionary selection arising from the division of labour between men and women in traditional hunter-gatherer societies (Eagly and Wood 1999; Geary 1995). Men would be required to travel long distances in order to track and hunt animals, a task requiring strong spatial perception and navigation skills (Buss 1995, 2015). In contrast, women fulfilled the role of the gatherer of more local food and assumed childrearing duties. This role had less need for spatial proficiency but emphasized other adaptive traits such as nurturing and fine-motor skills. Over successive generations, evolutionary forces may have developed sex-specific proficiencies in spatial ability, giving males a strong advantage over females with such tasks (Buss 2015; Jones et al. 2003).
Support for the position of evolutionary psychology comes from cross-cultural studies of cognitive gender differences. Unlike language and quantitative reasoning which shows substantial variation across countries and cultures (Else-Quest et al. 2010; Lynn and Mikk 2009; Reilly 2012), a large body of research has shown that spatial differences are consistently found in all countries (Janssen and Geiser 2012; Peters et al. 2006). Furthermore intelligence – including spatial ability – is a highly heritable trait (Bratko 1996; Sternberg 2012), meaning that it can be passed down from one generation to the next. Nevertheless, some researchers question the validity of evolutionary and genetic factors (Hyde 2014), arguing that at the genetic level men and women are identical with the exception of the sex chromosome. Such arguments do not take into account other biological differences. For instance, the expression of sex hormones might be an important factor linked to genetic and evolutionary gender differences (Hines 2015a; Sherry and Hampson 1997).

10.3.2 Contribution of Sex Hormones to Spatial Ability

Sex hormones such as androgens and estrogens have been proposed as a biological explanation for observed gender differences in spatial ability (Kimura 1996, 2000; Sherry and Hampson 1997). While both males and females produce these sex hormones to some degree, greater androgen production is typically found in males while greater estrogen and progesterone production is present in females. Such a difference starts early, with differences in testosterone concentration of foetuses found as early as 8 weeks gestation (Hines 2010). Production of sex hormones greatly increases with the onset of puberty (Spear 2000), and is associated with a range of psychological and behavioural changes as well as differences in brain development (Berenbaum and Beltz 2011; Sisk and Zehr 2005).

Even before birth, sex hormones contribute to the organisation and development of the brain with lasting effects on behaviour and interests for children (Hines 2015a). Girls exposed to higher than normal levels of androgenic hormones prenatally, either due to a genetic disorder such as congenital adrenal hyperplasia or because androgenic hormones were prescribed to mothers during pregnancy, show increased male-typical play, behaviour, and interests as young children (Auyeung et al. 2009; Hines 2010). Furthermore, they perform at a higher level on tasks of spatial ability than their same-sex peers (Puts et al. 2008). Because spatial ability requires environmental input for development, toys and play can be an important source of spatial experiences. Many stereotypically masculine activities such as construction blocks and model building promote spatial development (Caldera et al. 1989; Caplan and Caplan 1994), and gender differences in sex hormones may influence boys and girls play preferences.

Sex hormones also play an activational role in human behaviour and cognition after the onset of puberty (Berenbaum and Beltz 2011; Spear 2000), which coincides with a widening of the gender gap in spatial ability (Kimura 2000; Voyer et al. 1995). There is an intuitive appeal to considering hormones as explaining part or all
of the gender gap in spatial ability, but correlation by itself does not prove causation. Hormonal effects also coincides with increased gender conformity pressures for adolescents (Ruble et al. 2006) which may limit the interests and leisure activities that boys and girls pursue. These, in turn, may provide greater exposure to spatial experiences for boys than girls, thereby exacerbating gender differences.

To establish the causal effects of hormones would require an experiment whereby androgens were administered, which would be both impractical and unethical in developing children. There are instances where researchers have observed the effect of atypical levels of sex hormones (either reduced or increased levels) that are associated with certain medical conditions. Spatial ability in men diagnosed after puberty with hypogonadism is lower than in those with normal testosterone levels (Alexander et al. 1988; Hier and Crowley Jr. 1982), while men receiving hormone replacement therapy later in life showed significant improvements in spatial performance after treatment (Janowsky et al. 1994). In otherwise healthy individuals, some studies have also found a contribution of endogenous testosterone in the bloodstream to spatial performance in both genders (Davison and Susman 2001; Hausmann et al. 2009; Hromatko and Tadinac 2007), as well as fluctuations across the menstrual cycle in girls (Hausmann et al. 2000; Kimura and Hampson 1994). However, not every study finds robust associations (Puts et al. 2010), and the activation role that these hormones play may explain a much smaller proportion of variance in spatial ability than their earlier contribution to brain development (Falter et al. 2006).

10.3.3 Different Socialisation Experiences Between Boys and Girls

While biological contributions to spatial ability may explain some of the gender gap, many researchers argue that gender differences in early socialization experiences of boys and girls also play a significant role. Although there is certainly a contribution of biology, many theorists note that gender is socially constructed. From infancy and throughout childhood and adolescence, boys and girls experience the world differently, and are subject to different pressures and expectations (Lytton and Romney 1991; Martin and Ruble 2004). Boys and girls receive different messages about the suitability of particular toys from their parents, and elicit different styles of interaction during shared play with their parents, caregivers and siblings (Caldera et al. 1989). Children also acquire messages about gender expectations from their peers, and from their teachers and instructors once they have entered the educational system (Jacobs et al. 2002).

There are many different theoretical perspectives on the socialization of gender. For example, social-role theory proposes that psychological differences between men and women arise from gender segregation in men and women’s social roles (Eagly and Wood 1999), while the social cognitive theory of gender development posits that gender development is the result of learned experiences that teach gender roles through a system of observation, reinforcement, and punishment (Bussey and
An exhaustive coverage of the many other theoretical perspectives on gender is beyond the scope of this chapter, so we highlight only those relating specifically to spatial ability.

### 10.3.4 Sex-Role Mediation Theory of Spatial Ability

As children develop, they acquire stereotypically masculine or feminine traits, behaviours and interests, a developmental process referred to as sex-typing (Kohlberg and Ullian 1974; Martin and Ruble 2010). However, there is also wide variability across individuals in the degree to which people integrate masculine and feminine traits into their self-concept and sex-role identity (Bem 1981; Spence and Buckner 2000). Highly sex-typed individuals are motivated to keep their behaviour and self-concept consistent with traditionally gender norms, including the expression of intellectual abilities (Bem 1981; Steffens and Jelenec 2011). Others may integrate aspects of both masculine and feminine identification into their self-concept, termed androgyny.

The sex-role mediation hypothesis proposes that a masculine or androgynous sex-role identity promotes the development of spatial ability (Nash 1979). This theory proposes a number of mechanisms, including self-selection of play and leisure activities throughout childhood and adolescence, self-efficacy beliefs and motivation to practise tasks that encourage spatial competency, and sex-role conformity pressures (Reilly and Neumann 2013). This hypothesis has been tested a number of times over the decades, and two meta-analyses have been conducted (Reilly and Neumann 2013; Signorella and Jamison 1986). Both find support for sex-role mediation on the most prominently tested visual spatial task of mental rotation, but the scope of such reviews are limited by the shortage of studies testing other components of spatial ability. More recently an empirical study by Reilly, Neumann and Andrews (2016) tested support for the sex-role mediation hypothesis across a range of visual-spatial tasks, including mental rotation, spatial perception and spatial visualization. Masculine sex-role identification significantly predicted performance in both males and females.

### 10.3.5 Gender Stereotypes About Intelligence and Spatial Ability

Children begin to exhibit cultural stereotypes about what constitutes “masculine” or “feminine” by their early school years (Blakemore 2003; Ruble et al. 2006). This extends to characterising particular scholastic subjects and intellectual interests as masculine or feminine. For example, mathematics and geometry (which encourage development of spatial ability) is seen as masculine while language and arts are seen as feminine (Nosek et al. 2002). Boys also report greater interest and higher motivation in mathematics – a finding that is replicated cross-culturally (Goldman and Penner
Such stereotypes influence the way that men and women see themselves in relation to intellectual domains generally (Nosek et al. 2002), as well as their motivation to persevere when they encounter obstacles to learning (Meece et al. 2006).

While gender stereotypes may influence interest and motivation, they also shape perceptions of our abilities and self-efficacy. Despite there being no scientific evidence for gender differences in general intelligence, parents typically believe their sons are more intelligent than daughters (Furnham 2000; Furnham and Akande 2004; Furnham et al. 2002; Furnham and Thomas 2004). These gender stereotypes are quickly incorporated into children’s own self-beliefs and persist into adulthood. A consistent finding cross-culturally is that when asked to rate their own level of general intelligence, males tend to estimate their intelligence level considerably higher than do females (for a meta-analysis see Szymanowicz and Furnham 2011). The effect size of this gender difference is not insubstantial, $d=0.34$. Males also rate themselves as more spatially competent than females, $d=0.43$, which is again a moderately sized effect.

Popular cultural stereotypes (e.g. Pease and Pease 2001) that women can’t read maps or navigate without asking for directions do women a real disservice. Males in general are seen as more capable at performing spatial tasks by a significant degree (Halpern et al. 2011; Lunneborg 1982), and gender stereotypes can become self-fulfilling prophecies that undermine both interest in such tasks as well as performance (Steele 1997). Recognizing that spatial ability is not immutable, but that it can improve with learning and instruction is an important first step for any targeted intervention aimed at eliminating the gender gap and ensuring gender equity.

### 10.3.6 Differential Practice of Spatial Skills by Boys and Girls

Piaget (1951) was one of the earliest scholars to suggest that play is an important part of child development, helping to develop childrens’ motor skills and spatial abilities. Boys and girls are typically encouraged by parents to engage in stereotypically masculine and feminine play consistent with their gender (Eccles et al. 1990), but boys and girls also express preferences for different types of toys themselves (Hines 2015b). For example, boys tend to show a preference for vehicles and weapons while girls show more interest in dolls. The effect size for this gender difference is extremely large, with one study in children aged 4–10 years finding an effect size of $d=2.0$ (Pasterski et al. 2005). While there is considerable gender segregation in the types of toys marketed to boys and girls (Blakemore and Centers 2005), it is difficult to separate how much these choices are culturally directed and how much of the preference is biologically based. Recall that early androgen exposure prenatally has been associated with male-typical toy and play preferences (Auyeung et al. 2009; Hines 2010), suggesting at least some influence on boys’ and girls’ choices. Indeed, this strong effect is even found amongst non-human primates divorced of human cultural traditions. Male primates express greater interest and play longer with stereotypically masculine toys such as balls, cars, and trucks while female
primates preferred dolls and plush animals (Alexander and Hines 2002; Hassett et al. 2008).

Caplan and Caplan (1994) have argued that many stereotypically masculine toys and activities encourage the practice and development of spatial skills, while traditionally feminine play reinforces other culturally valued traits like communication and cooperation. For example, construction blocks and model assembly requires children to read 2D depictions of 3D objects and then find the correct spatial orientation of small and similar looking parts, while carpentry involves precise measurement of spatial relations and manipulation of parts. At earlier ages, toys like cars and trucks offer hands-on practice in visually tracking a moving object and judging the correct angle and speed to cause collisions. Girls play less on average with spatial toys than do males (Jirout and Newcombe 2015), and thus have less opportunities to practise these skills. Even if the effect of differential practice of spatial skills offers only a modest initial advantage to boys, the effect may grow larger as children enter adolescence and begin to self-select leisure activities and hobbies that they enjoy and are competent at performing. Activities such as carpentry, mechanics, models, and computer games would further enhance visual spatial skills.

There is strong evidence to support the theory that gender differences in spatial ability are at least partially influenced by differential levels of practice between boys and girls. Surveys and questionnaires measuring participation in spatial activities are positively correlated with performance on a range of spatial tests (Baenninger and Newcombe 1989; Chan 2007). However, it is equally plausible that people with high spatial ability may be the ones who want to engage in spatial activity in the first place (Baenninger and Newcombe 1989). It does seem likely that spatial activity experiences may be developmentally important in children (Doyle et al. 2012), and that differential levels of practice make some contribution.

10.4 Interventions for Training of Spatial Ability

A considerable body of evidence attests to the malleability of visuospatial reasoning, and that peak spatial ability is only reached with sufficient environmental input and experience (Baenninger and Newcombe 1995; Caplan and Caplan 1994). While biological and social factors may result in males starting with a modest initial advantage over females in spatial ability, it is important to remember that it is an acquired skill; people do not emerge de novo and become Tetris grand masters. There is an old joke that starts with the question “How do you get to Carnegie Hall?” – the punchline of course is “practice, practice, practice”. Like any other learned skill, if we receive training and do appropriate practice we can improve spatial abilities over time.

A large number of studies have examined the effects of brief training interventions to improve spatial ability. While there is wide variation in effectiveness, almost all such interventions show some improvement in spatial ability. With the large number of studies, training types, and choices of samples, the technique of meta-analysis
can provide an objective quantitative assessment. But before turning to these reviews, theoretical issues need to be considered.

There are four important theoretical questions. First, does spatial training benefit all recipients equally, or are there differential rates of improvement for males and females? If spatial training was only effective in those who already have a moderate level of proficiency, its usefulness in addressing the gender gap would be limited. Second, do the effects of training transfer to all spatial tasks (thereby indicating an improvement in latent spatial ability), or only to tasks that are very similar or indeed identical to those used in training? Sims and Mayer (2002) have questioned whether the effect of spatial training might simply be the result of practice and familiarity, rather than genuine improvement in latent ability. For interventions to be genuinely useful, training effects must generalise to novel and unfamiliar spatial tasks. Third, do the improvements to spatial ability persist over time or are they short-lived? Fourth, do all types of training interventions work, or do characteristics such as the type and intensity of training matter?

Two meta-analyses have investigated the effect of brief spatial instruction and training interventions. The first, by Baenninger and Newcombe (1989) investigated the effects of training in studies that used a repeated measures design (i.e. subjects’ initial performance on a spatial test is measured, a brief training intervention is offered, and then spatial performance is tested a second time). Their review included studies spanning a considerable range of years from the 1940s to the 1980s. They found that substantial improvements could be made to spatial ability after training, with an impressive effect size of \( d = 0.70 \) when tested on the same spatial measure that they were trained on, and a more modest effect size of \( d = 0.49 \) when more general spatial tasks were administered. This is an important distinction, because it shows that the effects of spatial training generalise well to other spatial tasks rather than being simply familiarity with the test content arising from repeated administration. The researchers also sought to test whether there was evidence of differential improvement between males and females, but found no significant gender differences. What the researchers did not address though is whether the improvements to spatial ability persist over time. Instead the authors considered the intensity of the training intervention, finding that multiple sessions over several weeks delivered meaningful improvement and that extremely brief or single session interventions showed less substantive benefits.

While the review by Baenninger and Newcombe (1989) makes an important contribution to the literature, a number of researchers have argued that changes in men and women’s roles over the past few decades should result in smaller gender difference over time (Caplan and Caplan 1994). When research becomes too dated, it raises the question of whether it remains applicable to current generations. More recently, Uttal et al. (2013b) conducted an extensive meta-analytic review of the empirical studies on spatial training from more recent years. Their meta-analysis also included a large number of unpublished studies (such as masters and PhD level theses). This is important because there might be a selection bias in the literature towards publishing only statistically significant findings while non-significant findings may be discarded, termed the file drawer effect in psychology (Ioannidis et al.
2014; Rosenthal 1979). A genuine test of the effectiveness of training interventions would also need to consider findings that might disconfirm the hypothesis.

Uttal et al. (2013b) considered a wide range of spatial training interventions, from explicit instruction and courses to playing video games and practising spatial tasks. The meta-analysis found that spatial training interventions were highly effective, with an overall effect size of $d=0.47$ which is a medium-sized effect. Consistent with the earlier meta-analysis by Baenninger and Newcombe there was no evidence for differential improvement between males and females. Both genders gained the same benefits from training. Moderator analysis also showed no difference in the type of training being offered, with similarly sized effects across interventions that offered spatial learning courses, practice on spatial tasks or practice on video games. Adults also showed similar rates of improvements as adolescents, and though there was a slight tendency for interventions with children to have larger effect sizes, this trend did not reach statistical significance.

Another important research question about training interventions is whether the effects persist over time. Most studies that report the results of a spatial training intervention test subjects at the conclusion of the intervention, but a number of the studies evaluated in Uttal et al. (2013b) introduced a short delay of a few weeks and some tested subjects after as long as several months (Terlecki et al. 2011). If there were genuine and lasting improvements to latent spatial ability, we should see similarly sized effects of improvement between studies that tested performance immediately to those studies that included some latency. The meta-analysis found the effect of training to be durable, with no diminution of improvement for studies that introduced a delay before retesting.

To address the question of whether training interventions show generalisability to other types of spatial tasks, Uttal et al. (2013b) compared studies that used very similar measures of spatial performance to that covered in training with studies that employed substantially different types of spatial tasks. Importantly, the meta-analysis showed no difference between these two categories, providing evidence of transfer to novel tasks.

The research outlined above provides strong evidence that regardless of gender, spatial ability is highly malleable with instruction and training. Furthermore these effects do transfer to other types of spatial tasks and persist over time. Even brief interventions seem to have some effect, but more intensive training over multiple sessions yields the strongest benefits. Importantly the effects of training generalise across tasks, and improvements can be delivered for practically any age group from children to older adults.

### 10.4.1 Spatial Training and STEM Outcomes

While spatial ability is important for many occupations, the most compelling benefits of spatial training are in improving mathematical and science achievement in students. Longitudinal studies have provided compelling evidence of an association
between spatial ability and proficiency in mathematics and science (Wai et al. 2010), but to date only a limited number of studies have investigated whether spatial training translates into tangible improvements in STEM achievement. Cheng and Mix (2014) conducted a randomized control trial of spatial training in a sample of 6- and 7-year old children, finding improvements in a test of basic calculation skills. A subsequent study by Krisztian et al. (2015) that taught spatial training with origami over a 10 week period in a sample of fifth and sixth grade students found similar improvements in computation skills over a control group. At present there are no spatial training studies that have measured science learning outcomes though in children, and none with adolescents in high school.

Amongst college-aged young adult samples, only two studies have investigated whether increasing spatial ability translates to improvements in mathematics and science learning. Sanchez (2012) conducted a randomized control trial that offered an intervention to target spatial ability, and found that the spatial group outperformed controls when tested on their learning from a short course on volcanoes and plate tectonics. In another study operating over a longer time period, Miller and Halpern (2013) recruited a sample of male and female first-year college students and randomly assigned them to either a control group or a spatial training condition (consisting of six 2-h spatial training sessions over a 6 week period). The gender gap in spatial ability narrowed somewhat after spatial training. In addition, the grades in student coursework were examined at the end of the year (up to 10 months after training ended). Compared to the control group, those receiving the intervention achieved higher grades in their physics coursework ($d=0.32$) but not in other classes like chemistry or calculus. The study also found significant correlations between students’ spatial ability and course GPA in the following sophomore year for a number of STEM courses, including electricity and magnetism, biology, engineering, and differential equations. The conclusions of this study are limited though by the small sample size for the treatment group (14 women, 24 men) which resulted in a reduced statistical power.

### 10.5 Reducing Gender Differences by Promoting Spatial Ability in Children

With the link between spatial ability and development of mathematics and science skills, a number of prominent educational and gender researchers have argued for the importance of developing spatial competency ability as a foundation for proficiency in STEM subjects (Hyde and Lindberg 2007; Newcombe and Frick 2010; Wai et al. 2009). With competing interests in a crowded curriculum, teachers and principals might be understandably reluctant to allocate time for regular lessons on promoting spatial competency. However, the effect of even brief training interventions over several sessions has been found to be effective in reducing the gender gap in spatial ability (Uttal et al. 2013a). Since both males and females can improve
their spatial reasoning substantially, it might be applied broadly to all students, which avoids the potentially stigmatizing effects of singling out females as a group for special interventions.

While explicit training would benefit older students such as those in high school or entering college, Newcombe and Frick (2010) advocate the importance of early education for spatial intelligence before the gender gap widens. One approach would be to integrate spatial learning with existing content in the STEM curriculum. In a report by the American National Research Council (2006), a range of practical strategies are outlined for engaging students to think spatially as part of mathematics and science classes. Rich multimedia can present complex scientific concepts visually, and many electronic textbooks offer data visualizations that are interactive rather than being static displays. For example, force and motion concepts are difficult to convey verbally or from a printed diagram. By showing the motion path of a physical object, a child can see the effects of physical phenomena.

Parents and caregivers might also gently encourage spatial learning outside of school by providing children with play and leisure activities (outlined in Table 10.1) that encourage spatial development through attention to spatial relationships (e.g., higher–lower; longer-shorter; wider-narrower). Games such as jigsaws, construction blocks, and board games provide contexts that facilitate spatial learning. Newcombe and Frick also note that everyday conversation can also be an opportunity

### Table 10.1 Summary of children’s play and leisure activities providing spatial experiences

<table>
<thead>
<tr>
<th>Age category</th>
<th>Play and leisure activity</th>
<th>Specific spatial abilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy and play experiences for younger children</td>
<td>Construction blocks</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>‘Action-oriented’ toys such as cars and vehicles</td>
<td>• •</td>
</tr>
<tr>
<td></td>
<td>Geometric shape toys</td>
<td>• •</td>
</tr>
<tr>
<td></td>
<td>Throwing and catching ball games</td>
<td>• •</td>
</tr>
<tr>
<td></td>
<td>Jigsaws</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>Art and drawing activities</td>
<td>• •</td>
</tr>
<tr>
<td></td>
<td>Mazes and maps</td>
<td>• •</td>
</tr>
<tr>
<td>Enrichment experiences for older children</td>
<td>‘Transforming’ toys appropriate to age</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>Advanced construction bricks such as Lego™</td>
<td>• •</td>
</tr>
<tr>
<td></td>
<td>Model building</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>Origami</td>
<td>• •</td>
</tr>
<tr>
<td></td>
<td>Computer games (action)</td>
<td>• • • •</td>
</tr>
<tr>
<td></td>
<td>Computer games (puzzle)</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>Computer games (construction)</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>Perceptual and motor skills training such as juggling</td>
<td>• • •</td>
</tr>
<tr>
<td></td>
<td>Organised sports</td>
<td>• • •</td>
</tr>
</tbody>
</table>

*SP* spatial perception, *MR* mental rotation, *SV* spatial visualization, *ST* spatiotemporal, *WF* wayfinding and navigation

---

10 Gender Differences in Spatial Ability
for parents to highlight the spatial properties of objects through questions and gently introduce spatial language and concepts into the conversation (Ferrara et al. 2011). Indeed, many household experiences can be learning opportunities to demonstrate spatial concepts, such as measuring and transformation of solids and liquids when moving ingredients from one container to another during cooking, or imagining what shape will be made if we fold a sheet of paper diagonally. Educational toys that provide examples of geometric shapes can be a good way to extend spatial language further by learning the names of common objects such as triangles, squares, circles, and relationships before introducing more complex shapes and concepts (Newcombe and Frick 2010).

Children as young as 3 or 4 years of age can understand the concepts of maps and how they relate to the physical world if introduced at the right pace (Shusterman et al. 2008), while puzzles like mazes can offer further practice of spatial and navigational skills (Jirout and Newcombe 2014). In older children, enrichment activities like jigsaw puzzles and origami can also provide additional opportunities to encourage spatial development (Boakes 2009; Taylor and Hutton 2013), particularly when parents and educators engage children in active conversation and provide guided assistance. Art and drawing activities can also provide practice in spatial perception and visualization skills (Calabrese and Marucci 2006). Age-appropriate toy robots that children can change into vehicles and back provides practice in learning complex multi-step transformations like that involved with spatial visualization, while a wealth of literature has shown that construction blocks provide opportunities to practise spatial perception and transformation skills (Caldera et al. 1999; Jirout and Newcombe 2015; Stannard et al. 2001). They also provide practice in interpreting two and three-dimensional diagrams, and then translating these diagrams into physical steps.

Another promising enrichment activity that aids in practising spatial skills may be video games. Computer gaming has emerged as a popular leisure activity for children and can be an opportunity to practise spatial skills. While boys still report playing more computer games than girls, in recent years the gap has been diminishing (Terlecki et al. 2011). Additionally, the wider availability of gaming on mobile phones and tablets may see shifts in gender patterns of usage. Not every player will enjoy first-person shooters or fast action games, and game developers are increasingly embracing other genres to entice non-game players into the market. However, not all games are equal, and some games may have greater educational potential than others. In a review by Spence and Feng (2010) on the contribution of video-game play to spatial cognition, action-based games and maze/puzzle genres emerged as the most likely to affect spatial cognition as they provide repeated practice in spatial perception, mental rotation, and navigation tasks. Indeed, a number of studies have shown that even brief training with computer games may be effective as an intervention (as reviewed earlier).

Parental concerns over the use of videogames may need to be considered if they are to be recommended. Concerns over violence in some types of videogames or excessive amounts of time spent playing remain legitimate (Festl et al. 2013). However, when enjoyed in moderation with parental selection of content there is evidence that the benefits for spatial cognition outweigh the costs (Ferguson 2007;
Parents may also be more comfortable offering less violent and adversarial games to their children, such as the popular construction and building game “Minecraft” which is appealing to boys and girls equally and is already used by some educators (e.g. Short 2012). Spence and Feng propose that gaming might also be an opportunity to deliver more targeted educational interventions specifically developed with the goal of raising spatial abilities in a similar fashion to commercial brain-training products.

There is also a strong link between the development of motor skills and spatial reasoning (Frick et al. 2009; Richter et al. 2000). Neuroimaging studies show that regions of the brain associated with motor skills are activated when performing mental rotation tasks (Halari et al. 2006; Richter et al. 2000). Interventions that consist of motor skills training have been shown to enhance mental rotation performance in children (Blüchel et al. 2013). Newcombe and Frick (2010) advocate that educators and parents should provide young children plenty of time for free play and physical action with objects like balls to provide practice in motor skills. By association, this should transfer into positive benefits for spatial ability.

Sporting activity and organised sports might also offer opportunities to more specifically develop spatial ability. While individual families may differ, sons typically receive greater encouragement to pursue athleticism and organised sports than daughters (Leaper 2005), and greater media attention and funding is given to male professional sports stars (Gill and Kamphoff 2010). In contrast, girls have lower enrolment in organised sports and withdraw from sporting teams at a higher rate (Vilhjalmsson and Kristjansdottir 2003). But there is evidence that playing sports may help to develop spatial ability (Moreau et al. 2015). When children who play regular sport were compared to similar aged matches who did not, those who played sport performed better on tests of spatial performance (Notarnicola et al. 2014), with similar findings in young adults (Lord and Leonard 1997; Moreau et al. 2011).

Motor coordination is a significant predictor of mental rotation ability even after controlling for the effect of gender (Pietsch and Jansen 2012), and two studies have found that learning and practising juggling skills increased mental rotation performance for both adults and children (Jansen et al. 2009, 2011). Encouragement of sports activity within the context of the educational system and by parents may help to lessen the gender gap in spatial ability, in addition to the non-cognitive benefits (Moreau et al. 2015).

### 10.6 Directions for Future Research

Most researchers now endorse biopsychosocial models of gender differences in spatial ability (Halpern et al. 2007) rather than considering exclusively biological or social causes, and the debate has shifted towards their relative contributions. Whereas once spatial ability was considered fixed and immutable, a considerable body of research has demonstrated that exposure to new spatial experiences throughout early childhood promotes growth in spatial proficiency. Furthermore, spatial training interventions can produce substantial benefits that potentially could...
translate to a reduction or even the elimination of the gender gap in mathematics and science achievement.

As reviewed earlier, only a limited number of spatial training studies have measured subsequent outcomes in science and mathematics achievement outcomes however. To date though, there have been no spatial training interventions that have followed children longitudinally to follow their progress, and only a single study by Miller and Halpern (2013) has tracked the progress of college-aged students for a prolonged length of time. Arguments for spatial training interventions would be strengthened by further studies monitoring student progress over longer time periods. It would also allow investigators to determine what types of spatial training and at what intervals, will best deliver changes in STEM-specific outcomes. While brief interventions may well yield long-term improvement, it is also possible that spatial training will require maintenance “booster” training at periodic intervals to deliver lasting educational improvements.

10.7 Summary and Conclusions

While individuals may differ, on average males score higher in tests of visual spatial ability. They also rate themselves as more spatially competent than females. Gender differences in spatial ability emerge from an early age. While clearly observable in children, the gender gap widens in adolescence and continues to grow into adulthood where it is quite large. Gender differences are found for a variety of categories of spatial tasks, but the largest and most actively studied is mental rotation, followed by spatial perception and then spatial visualization skills. There are a range of theoretical perspectives on why gender differences in spatial ability develop from biology to environmental causes, but one of the most frequently argued causes is differential levels of spatial learning and practice between males and females. This is supported by retrospective studies finding associations between childhood spatial experiences and spatial ability in adults.

Gender differences in spatial ability also precede the development of gender differences in mathematics and science, and longitudinal studies have found that early performance on spatial tasks can predict future performance in STEM, even many years later. There is also robust evidence demonstrating that spatial ability is not an immutable skill, and that even brief interventions can deliver impressively sized improvements. Such evidence makes a compelling argument for integrating spatial learning into early education, but parents can also provide additional learning opportunities for their children by engaging in spatial language, demonstrating spatial concepts within the home, and providing toys and games that encourage spatial practice. In older children, computer games can provide an opportunity to learn and practise spatial skills if they express an interest them, and organised sports has also been shown to improve spatial ability. The research supports the conclusion that concerted efforts by educators to address the gender gap in spatial ability in children and adolescents may translate into improvements in girls’ and boys’ mathematics
and science achievement. However there is a need for longitudinal studies to determine which types of training and at what intervals will best support students in this regard, and the extent to which this reduces the gender gap for STEM outcomes.

References


Gender Differences in Spatial Ability


Sex Differences in Mathematics and Science Achievement: A Meta-Analysis of National Assessment of Educational Progress Assessments

David Reilly, David L. Neumann, and Glenda Andrews


CITATION
Sex Differences in Mathematics and Science Achievement: A Meta-Analysis of National Assessment of Educational Progress Assessments

David Reilly, David L. Neumann, and Glenda Andrews
Griffith University

Gender gaps in the development of mathematical and scientific literacy have important implications for the general public’s understanding of scientific issues and for the underrepresentation of women in science, technology, engineering, and math. We subjected data from the National Assessment of Educational Progress to a meta-analysis to examine whether there were sex differences in mathematics and science achievement for students in the United States across the period 1990–2011. Results show that there were small but stable mean sex differences favoring males in mathematics and science across the past 2 decades, with an effect size of $d = .10$ and .13, respectively, for students in 12th grade. Furthermore, there were large sex differences in high achievers, with males being overrepresented by a factor of over 2:1 at the upper right of the ability distribution for both mathematics and science. Further efforts are called for to reach equity in mathematics and science educational outcomes for all students.

Keywords: sex differences, mathematics, science, education, meta-analysis

The issue of sex differences in science and mathematics achievement continues to capture the interest of parents, educators, researchers, and policy makers and has implications for the ways in which children are educated and encouraged to pursue their chosen careers (Halpern et al., 2007; Hyde & Lindberg, 2007). Although significant inroads have been made in recent decades, women continue to be underrepresented in fields related to science, technology, engineering, and math (STEM; Handelsman et al., 2005; National Science Foundation, 2011), even though more women than men now attend college (Alon & Gelbiser, 2011). Predicted shortfalls in the number of science graduates for the United States relative to other developing nations carry serious economic and social consequences (President’s Council of Advisors on Science and Technology, 2010) and will require broadening the pool of new entrants into STEM fields to include more women in order to meet the growing demand. Though the exact causal mechanisms that contribute to sex differences in entering mathematics and science fields are yet to be fully understood (Ceci & Williams, 2011; Hanson, Schaub, & Baker, 1996), many researchers believe that early sex differences in achievement at school shape attitudes toward STEM fields and self-efficacy beliefs (Halpern et al., 2007; Newcombe et al., 2009; Wai, Lubinski, & Benbow, 2009; Wang, Eccles, & Kenny, 2013). Furthermore, even if they choose not to pursue a STEM-related profession, students entering college and university are increasingly required to have more advanced technical and quantitative skills. For this reason the emergence of sex differences in educational achievement of students is of interest to educational psychologists.

A key component of any strategy to raise the representation of women in STEM fields is to address gender gaps in mathematics and science outcomes, but the existence and magnitude of these differences are strongly contested (Gallagher & Kaufman, 2005; Halpern et al., 2007; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 2006; Spelke, 2005; Wai et al., 2009). Much of the empirical research in this area is somewhat dated (e.g., Hyde et al., 1990). Furthermore, as Hedges and Nowell (1995) pointed out, with few exceptions most empirical studies in this area are subject to selection and sampling biases. Furthermore, as there are interactions between gender and other sociocultural factors (Becker & Hedges, 1988; Frieze, 2014; Hyde & Mertz, 2009; Nowell & Hedges, 1998; Spelke, 2005) these findings do not necessarily generalize well to the wider population. Debate about educational issues such as sex-segregated schooling (Halpern, Eliot, et al., 2011) or early intervention programs to boost mathematics and science literacy (Hyde & Lindberg, 2007; Newcombe & Frick, 2010) can only be served by timely and accurate empirical research into the nature of sex differences in science and mathematics achievement (Alberts, 2010; Halpern, Beninger, & Straight, 2011). Additionally, if gender gaps are decreasing in response to cultural and educational changes (Auster & Ohm, 2000; Wood & Eagly, 2012), existing research on sex differences in educational achievement for mathematics and science could quickly become dated and require periodic reassessment (Hyde & Mertz, 2009). We describe the findings of prior research on sex differences in these domains and then extend these findings by reporting a meta-analysis of sex differences in national science and mathematical achievement from the National Assessment of Educational Progress (NAEP) for
the years 1990–2011. First, we review the theoretical frameworks that posit the emergence of sex differences in quantitative reasoning.

**Theoretical Perspectives on Sex Differences in Quantitative Reasoning**

Although reviews of intelligence testing studies find no evidence for sex differences in general intelligence (Halpern & Lamay, 2000; Neisser et al., 1996), consistent patterns of sex differences have been observed for more specific components of cognitive ability (Halpern, 2011; Kimura, 2000). For example, women show greater proficiency with verbal ability and language tasks and men demonstrate higher performance on tasks that tap visuospatial abilities (Halpern & Lamay, 2000). Sex differences have also been documented in quantitative reasoning (our present focus) in tasks that assess mathematical and scientific skills (Halpern et al., 2007; Wai et al., 2009). A number of theoretical perspectives have been proposed by researchers to explain why sex differences in quantitative reasoning might emerge; these include both biological and psychosocial contributions. Although a full critique of all these theoretical perspectives is beyond the scope of this study, the most prominent and well-established perspectives may be categorized as biological, social/environmental, or psychobiosocial theories.

**Biological Theories of Sex Differences**

Sex hormones have been proposed as an explanation for group differences between males and females (Collins & Kimura, 1997; Kimura, 2000), because sex hormones exert an influence on the organization and development of the human brain before birth (Hines, 2006) and play an activational role at different points in maturation (Hines, 1990). Associations have been found between digit ratio—a marker of prenatal androgen exposure—and some cognitive tasks (Collaer, Reimers, & Manning, 2007), though evidence has been mixed. However, most research on biological contributions to sex differences has focused on sex hormone production, which increases with the onset of puberty. Because this increase coincides with a widening of the gender gap in quantitative reasoning during adolescence and early adulthood (Hyde et al., 1990), there is an intuitive appeal to such an explanation. Although initial interest by researchers into the contributions of sex hormones such as androgens to sex differences in quantitative reasoning was high (Kimura & Hampson, 1994), research findings have found mixed support. Some studies have found no association, and other studies have observed that endogenous hormone levels explain very little variance in individual performance (Halari et al., 2005; Puts et al., 2010).

Another purported biological contribution to sex differences in quantitative reasoning comes from evolutionary psychology. Darwin (1871) first proposed that sexual selection as a result of evolutionary pressures has led to a differentiation in the roles of men and women, a theme that has been expanded upon by evolutionary psychology to propose an alternate explanation for why sex differences in quantitative reasoning emerge (Archer, 1996; Geary, 1996). In the past, it was adaptive for males to develop and hone spatial skills for navigation and hunting (Buss, 1995), leading to the development of greater visuospatial ability in males. This in turn lays down the foundation for the development of quantitative reasoning through a variety of mechanisms including differing social roles and sex typing of children’s’ play activities (Caplan & Caplan, 1994; Geary, 1996, 2010). Furthermore, the traditionally feminine roles of caring for others and sensitivity to emotions may have been adaptive, resulting in a tendency for women to focus on people over things (Su, Rounds, & Armstrong, 2009), which Hyde (2014) argued may decrease motivation to acquire quantitative skills and pursue a STEM-based career. A common theme in such arguments is an interaction between biology and environment, rather than a strictly deterministic role of biology.

**Social and Environmental Contributions**

Although biological factors may make a modest contribution to sex differences, many theorists argue that psychological and social factors exert a greater influence over the course of a lifetime. One such theory is Eagly and Wood’s social-role theory (Eagly, 1987; Eagly & Wood, 1999), which proposes that any psychological sex differences arise from the distribution of men and women’s roles in society. The gendered division of labor between men and women encourages the development of instrumental and achievement-oriented traits in men and expressive and communal-oriented traits in women. Such a position is also compatible with gender schema theory (Bem, 1981), which proposes that children develop an internal schema about the sex typing of interests and behavior and that they are motivated to behave in a manner consistent with their internal sex-role identity (Martin & Ruble, 2004). From an early age children learn to categorize things as inherently masculine or feminine (Kagan, 1964), including school subjects like mathematics and science (Nosek et al., 2009). These form the foundation for sex typing of interests and activities, which facilitates the development of specific cognitive abilities. Nash (1979) formalized this as a sex-role mediation explanation for cognitive sex differences, theorizing that masculine identification leads to cultivation of spatial, mathematical, and scientific skills (Reilly & Neumann, 2013; Signorella & Janison, 1986).

Another prominent theory was put forward by Caplan and Caplan (1994), who argued that traditionally “masculine” play activities promote the development of spatial ability by encouraging the practice and application of spatial skills (Serbin, Zellkowitz, Doyle, Gold, & Weathen, 1990). Other theorists argue that gender conformity pressures also play an affective role in developing one’s talents. Highly sex-typed individuals are motivated to keep their behavior consistent with internalized sex-role standards and norms, but those low in sex typing show greater cognitive and behavioral flexibility (Bem, 1975; Martin & Ruble, 2004; Spence, 1984). This has implications for success in academic domains that are traditionally male dominated, such as science and mathematics (Eccles, 2007). Conversely, as we see changes in the segregation of men’s and women’s roles and increasing gender equality, we might also see a diminishing of sex differences in these areas over time (Hyde, 2014).

**Psychobiosocial Theories of Sex Differences**

Theorists may be divided over the relative share of nature and nurture in the emergence of sex differences in cognitive abilities, but there is a growing consensus that both make a meaningful
contribution and neither in isolation cannot explain sex differences (Wood & Eagly, 2013). Indeed, it may be impractical to separate a specific biological and social component and study them in isolation, as their effects are reciprocal in nature (Halpern, 2011). Many theorists have therefore adopted psychobiological models for explaining the development of sex differences (Halpern & Tan, 2001; Hausmann, Schoofs, Rosenthal, & Jordan, 2009); these incorporate elements of biological, psychosocial, and sociocultural factors to explain group differences between males and females at the population level.

These theories offer perspectives on why sex differences in quantitative reasoning may be found, but it is also important to consider the many ways in which males and females are alike. Hyde (2005) has proposed the gender similarities hypothesis, which argues that men and women are more similar than different. Specifically, it hypothesizes that sex differences in cognition are either small in magnitude or nonexistent. Although this hypothesis is not supported for language (Lynn & Mikl, 2009; Stoet & Geary, 2013) and spatial abilities (Voyer, Voyer, & Bryden, 1995), where sex differences are moderately large, the gender similarities hypothesis may be compatible with the existence of sex differences in quantitative reasoning, as these tend to be somewhat smaller in magnitude (Hyde et al., 1990). However, the gender similarities hypothesis would be incompatible with sex differences that are moderate or large in magnitude, such as a gender imbalance in the sex ratio of high-achieving students in mathematics and science (Benbow, 1988; Hedges & Nowell, 1995). It is also a hypothesis that is can easily be put to the test, by examining the performance of men and women in tests that tap quantitative reasoning skills.

### Previous Meta-Analyses of Sex Differences in Mathematics and Science

Meta-analysis of national testing data by Hedges and Nowell (1995) from several decades of assessment (1960s–1990s) revealed small mean differences favoring males in mathematics and science performance (ranging from $d = .03$ to $d = .26$ for mathematics and $d = .11$ to $d = .50$ for science). Although mean sex differences might play an important role in the underrepresentation of women in STEM fields, other researchers have noted that the distribution of performance in a number of cognitive domains is more variable for males than for females (Feingold, 1992; Hyde, 2005; Machin & Pekkarinen, 2008). Even if there were no differences in the average performance of males and females on a specific ability test, greater variance in the male group would result in an overrepresentation in the extreme tails of the distribution (Feingold, 1992; Halpern et al., 2007; Turkheimer & Halpern, 2009), such as the intellectually gifted from which many STEM researchers hail (Wai, Cacchio, Putilaz, & Makel, 2010). For example, sex (male:female) ratios of students at the 95th percentile in the above-mentioned data sets ranged from 1.5 to 2.4 in mathematics and 2.5 to 7.0 in science achievement across samples (Hedges & Nowell, 1995). This can translate to a disparity in educational outcomes, and some researchers have argued that sex differences in variability may be more important than the mean differences (Feingold, 1995; Humphreys, 1988; Machin & Pekkarinen, 2008).

The greater male variability hypothesis can be examined through calculation of the variance ratio (VR), defined as the ratio of male variance to female variance (Feingold, 1992; Hedges & Nowell, 1995; Turkheimer & Halpern, 2009). A variability ratio of 1.00 indicates that males and females are equal in variance. VR values less than 1.00 indicate that females show more variability than males, and VR values greater than 1.00 reflect greater male variability (Priess & Hyde, 2010). Feingold (1994) argued that values between 0.90 and 1.10 ought to be regarded as negligible (i.e., homogeneity of variance), and this practice is adopted herein.

More recently, Hyde, Lindberg, Linn, Ellis, and Williams (2008) presented data from a subset of the National Assessment of Educational Progress (NAEP), a nationally representative probability sample drawn from all 50 U.S. states. The advantage of this sampling method is that national NAEP data provide a reliable population-level estimate of student performance, reflecting the demographic traits of the general population of students. Although individual state and national performance data were not available at the time, Hyde et al. (2008) obtained data from a selection of 10 states across Grades 2 through 11. Mean sex differences were small ($ds$ from $.02$ to $.06$). Hyde (2014) has characterized these differences as “trivial” in size, and others have used this research to argue that sex differences are no longer found in modern samples (Hyde & Mertz, 2009; Lindberg, Hyde, Petersen, & Linn, 2010).

Although Hyde et al. (2008) conducted their analysis with the most recent information available at the time, a key limitation of their methodology is that only a 10-state subset of the national data set was analyzed. Hedges and Nowell (1995) argued there are limitations to the use of samples that show a selection bias, because the conclusions they yield may be erroneous if attempting to generalize to the wider population (Becker & Hedges, 1988; Spelke, 2005; Stumpf, 1995). In particular, use of such samples may affect the magnitude of any observed gender gap, as literature suggests an interaction between student and socioeconomic background for many cognitive abilities (Hanscombe et al., 2012; Levine, Vasilyeva, Lourenco, Newcombe, & Huttenlocher, 2005). National assessments of the NAEP are also drawn from both public and private schools and thus may better reflect the demographic composition of students enrolled in U.S. educational institutions than analysis of only public school data.

The national test data from the NAEP are now publicly available for researchers, and they provide a broader sampling of students than was available at the time to Hyde et al. (2008). We present an analysis of national NAEP performance for boys and girls, allowing for an empirical test of claims of sex differences in mathematics for U.S. students in the present day. Furthermore, because data are now available across several decades, it is possible to examine temporal trends across the year of assessment as well as developmental trends across grade level of students (Hyde et al., 2008; Lindberg et al., 2010). Although the NAEP assesses mathematics more regularly, periodic national testing of science performance makes it possible to assess gender gaps in this domain as well. Sex differences in science achievement may also play a role in the decision of individuals to pursue a science-related profession.

We focused on four key research questions for the domains of mathematics and science. First, are there sex differences in overall mathematics and science achievement for modern samples of students in the United States, and is the gap diminishing over time? Sociocultural theories of sex differences would predict a decline in the magnitude of sex differences over time, but biological and
psychobiosocial theories would be compatible with stability in effect sizes.

Second, do males show greater variability in performance than females, as predicted by biological theories? Third, if there are sex differences in means and in variance, what is their combined contribution to the proportion of males and females attaining an advanced proficiency standard in mathematics and science achievement? Finally, if there are sex differences in science achievement, are they present for all of the three content areas assessed (earth science, physical sciences, life sciences)? These research questions also provide a test of the sex differences and similarities hypothesis, which would predict that effect sizes are small in magnitude.

Method

National Assessment of Educational Progress Data Source

The NAEP is a project of the National Center for Education Statistics (NCES), part of the U.S. Department of Education. NAEP conducts assessments across a range of subjects, including reading, writing, mathematics, history, civics, geography, and science. Each subject area is assessed periodically, and the most frequently assessed subjects are reading, mathematics, and science. National and state performance in each assessment are reported publicly in a series of documents titled “The Nation’s Report Card.” These documents provide a review of major trends written in language accessible to parents, educators, and policy makers (http://nces.ed.gov/nationsreportcard.gov/). These form part of the main NAEP assessment, which uses a modern mathematics and science curriculum with large sample sizes and frequent assessments. A secondary category of assessment is the NAEP long-term trends (LTT) assessment of mathematics, which samples students on an earlier curriculum framework from the 1970s onward. The LTT assesses more basic mathematical content, such as numbers, shapes, measurement, and probability; the main assessment also includes algebra, geometry, and problem solving. Additionally, the LTT restricts students to hand calculations, which limits the depth of complexity for assessment items. Although useful information can be obtained from the long-term trend assessments, it fails to adequately assess students’ knowledge of more advanced mathematical content included in the main assessment frameworks and is sampled less frequently than the main assessment. As such, it was deemed unsuitable for analysis, and only the main assessment data were reported in the main article. However, published reports of the LTT long-term assessments show a consistent gender gap in favor of males in mathematics for students at age 13 and 17 that has remained essentially unchanged since assessments began (Rampey, Dion, & Donahue, 2009).

The results of NAEP assessments are made freely available to researchers for secondary analysis via the NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepadata/). The target population for NAEP national assessments is made up of all students in any educational institution (from both private and public schooling), currently enrolled in the target grade (4, 8, and 12). School and student responses are appropriately weighted to draw an estimate of the target population that reflects student demographics (e.g., specific ethnic and socioeconomic groups). This may mean that some students and schools will be oversampled or undersampled, as appropriate. These weights are applied to draw an estimate of national student performance, reported through the NAEP Data Explorer. Additional information about sampling design is available from the NAEP website (https://nces.ed.gov/nationsreportcard/mathematics/sampledesign.asp).

Mathematics framework. The mathematics assessment framework covers five key content areas, which have remained the same since 1990. These are (a) number properties and operations; (b) measurement; (c) geometry; (d) data analysis, statistics, and probability; and (e) algebra. Students are assessed at a grade-level-appropriate standard (for example, at Grade 8 the topic of algebra includes linear equations, whereas at Grade 12 this topic is extended to include quadratic and exponential equations). Assessment items vary in complexity level to accommodate a wide range of ability levels. This is important, as some research has noted greater sex differences are present for complex problem-solving items (Hyde et al., 1990). Calculators are permitted for approximately one third of the assessment, but the remaining questions must be completed without calculators. The mathematics framework for assessment of Grades 4 and 8 is comparable with that for earlier assessments, allowing student performance in more recent years to be compared to those from earlier assessments. Although a revised mathematics framework was instituted in 2005 for students in Grade 12, these assessments are comparable to those administered previously, as they reflect similar content areas. Further information on the mathematics content areas can be found at the NAEP website (http://nces.ed.gov/nationsreportcard/mathematics/whatmeasure.aspx).

Science framework. Topic areas for science assessment are grouped into the following three domains, which both form separate subscales and contribute to the overall science achievement score:

- Physical sciences, including concepts related to properties and changes of matter, forms of energy, energy transfer and conservation, position and motion of objects, and forces affecting motion.
- Life sciences, including organization and development of cells and organisms, matter and energy transformations, interdependence, heredity and reproduction, evolution and diversity.
- Earth and space sciences, including concepts relating to objects in the universe, the history of the Earth, material properties, tectonics and energy in Earth systems, climate and weather, and biogeochemical cycles.

The science framework used for assessment was revised in 2009, in response to revised national science education standards. Although the content areas remained the same (physical, earth, and life science), they now include coverage of space science. Students completed a range of multiple-choice and open-ended questions, including hands-on practical science tasks and interactive computer-administered tasks, from the 2009 assessment onward. For additional information about the science framework and sample questions, see http://nces.ed.gov/nationsreportcard/science/whatmeasure.aspx.

Reliability of the NAEP instrument. Multiple choice items are computer scored, and constructed response are marked by raters. Consistency across markers for the constructed response items was generally high for both mathematics and science (Cohen’s κ > .80). Item response theory is then employed by NCES to measure latent scores, which offers greater control over the
measurement characteristics of each question and ensures high reliability. (See http://nces.ed.gov/nationsreportcard/tlw/analysis/ for additional information about reliability of measures.) Furthermore, the NCES conducted a NAEP–Trends in International Mathematics and Science Study (TIMSS) linking study to compare the assessment frameworks to international standards, finding them comparable.

Schedule of Assessment

Mathematics and science assessments are conducted periodically, in adherence with the NAEP schedule. Mathematics is assessed more frequently, roughly every two to three years (1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011). The schedule of assessments gives greater coverage to Grades 4 and 8, which are developmentally critical time periods for the acquisition of mathematics and scientific skills (Newcombe & Frick, 2010). Grade 12 assessment was not conducted in 2003 and 2011. Science is assessed every four to five years (1996, 2000, 2005, 2009, 2011) and with somewhat smaller samples of students than for the mathematics assessments. Grades 4, 8, and 12 were all assessed in the science target years, except for 2011.

In addition to achieving an overall test score, students are evaluated against fixed achievement levels in the NAEP, which categorizes students at a basic, proficient, and advanced level. Sex differences in the percentage of students attaining these levels are also available and were obtained from the NAEP Data Explorer. Although some researchers have examined sex differences in the extreme upper tail of mathematics and science distributions (Benbow, 1988; Hedges & Nowell, 1995; Hyde et al., 2008; Nowell & Hedges, 1998; Wai et al., 2010), Hyde and Mertz (2009) have questioned whether sex differences in extreme talent are a necessary requirement for pursuing STEM-related fields. When greater male variability is present, this may present an exaggerated picture of sex differences, particularly if more stringent cutoff points are examined (e.g., 99.9th percentile). Examining sex ratios in attainment of an advanced proficiency in science or mathematics represents a trade-off between selecting a cutoff point that is germane to the question of underrepresentation of women in STEM-related fields and seeking to avoid selecting an ability level that serves to exaggerate sex differences.

Participants

National performance data in NAEP mathematics were examined for the period 1990–2011, with a combined total sample size of almost 2 million students (see Table 1). Performance data in science were examined for the period 1996–2011. Science was assessed less frequently and with fewer students, with a combined total sample size of over 800,000. Information on sample sizes was obtained from annual reports of the NAEP, which in recent years followed the convention of rounding to the nearest hundred. When individual numbers of males and females were not reported, the assumption of equal sample sizes was made. Additional information on the schedule of assessments and sample size of individual assessment years can be found in the Appendix.

<table>
<thead>
<tr>
<th>Content domain</th>
<th>Grade</th>
<th>N students assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>4</td>
<td>974,700</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>845,400</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>104,900</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,925,100</td>
</tr>
<tr>
<td>Science</td>
<td>4</td>
<td>352,105</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>470,374</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>56,437</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>878,916</td>
</tr>
</tbody>
</table>

Meta-Analytic Procedure

Mean math and science scores and standard deviation for males and females were obtained from the Data Explorer website. The NAEP Data Explorer provides summary statistics (i.e., mean, standard deviation) rounded off to whole numbers, which introduces measurement imprecision. It can also export more precise values in Excel format, which was the option used in this meta-analysis. The unit of analysis was group differences in performance of males and females at the national level, rather than for individual states. Effect sizes are reported as the mean difference between males and females in standardized units (Cohen, 1988; Hedges, 2008), commonly referred to as Cohen’s d. By convention, a positive value for d indicates higher male performance and a negative value indicates higher female performance (Hyde, 2005).

Comprehensive Meta Analysis (CMA) V2 and Microsoft Excel software were used to calculate the statistics. Meta-analysis typically employs either a fixed-effects or a random-effects model for combining study samples. As NAEP assessments span a number of decades recruiting from independent samples, and it was hypothesized that student characteristics may have changed across years of sampling, a random-effects model was chosen (Borenstein, Hedges, Higgins, & Rothstein, 2009). The random effects model gives slightly wider confidence intervals than a fixed-effects model, but it gives a more appropriate estimate of how much variability is present across samples (Hunter & Schmidt, 2000; Kelley & Kelley, 2012). The benefit of such an approach is that we can have greater confidence in the population estimate of sex differences produced and that it is not the result of inflated Type I error. Using a random effects statistical model also adjusts for variation in test content and student characteristics over time.

In addition to calculating effect size data for each grade level, we investigated whether the year of assessment was a potential moderator with the technique of meta-regression (Kelley & Kelley, 2012). Meta-regression extends a conventional meta-analysis by determining whether a moderating variable accounts for variation in the magnitude of an observed effect (i.e., explains sources of heterogeneity). Based on claims of diminishing gender gaps (e.g., Hyde & Linn, 2006), a negative association with year of assessment was predicted. Although it is clear that sex differences in mathematics are smaller than systematic reviews had found in data from the 1960s–1980s (Hedges & Nowell, 1995), it is not apparent whether such a trend would continue to the point at which males

Table 1

Sample Size Information for Mathematics and Science Assessments

<table>
<thead>
<tr>
<th>Content domain</th>
<th>Grade</th>
<th>N students assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>4</td>
<td>974,700</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>845,400</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>104,900</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,925,100</td>
</tr>
<tr>
<td>Science</td>
<td>4</td>
<td>352,105</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>470,374</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>56,437</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>878,916</td>
</tr>
</tbody>
</table>
and females would perform equivalently (Caplan & Caplan, 1994) or whether it would plateau. We employed a random effects model (method of moments) for the meta-regression model to test if the year of assessment acted as a moderator (Borenstein et al., 2009; Thompson & Higgins, 2002). Additionally, we performed subgroup analysis for individual grades with a random effects model to examine whether sex differences change as students progress through their schooling, as indicated by previous research (Hyde et al., 1990).

Variance ratios (VR) for individual samples were calculated following the method of Feingold (1992). Estimates of overall male and female variance ratios were combined across years of sampling for each grade level. Some researchers have questioned whether, when variance ratios are combined across samples, mean variance ratios may be the most appropriate measure (Katzman & Alliger, 1992) and have advocated the use of medians or log transformed means. These metrics are most appropriate if the direction of variance ratios change across samples (i.e., greater male variability is found in some samples, and greater female variability is found in others). Although this was not the case (see the Appendix), by convention and for comparability with other studies the log transformed variance ratios were averaged across sample years and then transformed back into the Fisher’s variance ratio statistic. This statistic addresses whether males and females differ at the extreme tails of an ability distribution (e.g., the top 1% of gifted students) rather than focusing on the performance of the “average” students in the middle of the distribution (Priess & Hyde, 2010).

Additionally, the percentages of students for each gender who achieved an advanced proficiency standard were obtained to investigate the combined effect of sex differences in central tendency and variability. Sex ratios, defined as the relative risk ratio (RR) of male to female students, were calculated for mathematics and science performance at the advanced level of proficiency. This methodology is somewhat different than that followed in previous studies. It represents a trade-off between selecting a cutoff point that fairly evaluates high-achieving students in their ability to solve STEM problems and selecting an arbitrarily high cutoff (e.g., 99th percentile) that would serve to exaggerate sex differences.

Results

We conducted two separate meta-analyses on the NAEP sample for mathematics and science, with population-level estimates of sex differences partitioned by grade level (4, 8, 12). Although statistically significant sex differences favoring males were found in each grade ($p < .001$), emphasis is placed on effect size, as this gives an indication of the magnitude and practical impact of the observed differences (Hedges, 2008; Hyde, 2005). In a review of meta-analytic theory and practice, Hyde and Grabe (2008, p. 170) recommended a threshold for considering effect sizes in sex differences research a priori and argued that effect sizes smaller than $d = .10$ be considered “trivial” per Hyde’s (2005) gender similarities hypothesis. Accordingly we use this threshold herein for considering whether the observed sex differences are practically meaningful. Variance ratios and the sex ratio of students attaining the advanced level of proficiency are also reported for math and science. The original data used in this analysis are presented in the Appendix.

NAEP Assessment of Mathematics

National performance data in mathematics were examined for the period 1990–2011 (see the Appendix for a schedule of assessment years). National sex differences are somewhat larger than those reported by Hyde et al. (2008) in their 10-state sample, with a weighted mean effect size of $d = .07$, $Z = 12.07$, $p < .001$. However there was considerable heterogeneity present in the distribution of effect sizes, $Q(23) = 251.57$, $p < .001$, $F = 90.86$ (see Figure 1). In order to better explain variability across assessments, we tested whether grade level and year of assessment were potential moderators.

Grade level as a moderator. Table 2 presents comparisons between males and females in math across the three grade levels. When effect sizes were partitioned across the three measured age groups with subgroup analysis, there was a statistically significant difference between grade levels, $Q(2) = 23.15$, $p < .001$. Although sex differences were extremely small in elementary and early high school, they grew larger in the final year of high school ($d = .10$). The Grade 12 effect size is at the threshold of Hyde’s (2005) criterion for nontrivial sex differences.

Year of assessment as a moderator. Next we performed a meta-regression analysis to test for a declining gender gap in mathematics over time. Contrary to our hypothesis, there was no significant effect of assessment year, $Z = -1.10$, $b = -.0001$, $CI_{.95} [-.0016, .0015]$, $p = .923$; nor was the interaction between year and grade significant. This is consistent with other studies that reported stability for mean sex differences in mathematics in recent decades rather than a declining trend (McGraw, Lubienski, & Strutchens, 2006; Rampey et al., 2009).

Variance ratios. In line with previous research, the variability of males’ performance in mathematics was wider than that of females across each age group (see Table 3) and exceeded Feingold’s (1994) threshold for nontrivial variance ratios. These variance ratios were also stable across the time period examined, with no association with year of assessment or grade ($p > .05$).

![Figure 1. Histogram of observed effect sizes in NAEP mathematics assessments (1990–2011). NAEP = National Assessment of Educational Progress.](image-url)
Gender gaps in high achievers for mathematics. In order to evaluate the combined effect of mean differences and greater male variability, we calculated the ratio of males:females attaining the advanced proficiency standard for mathematics, \( RR = 1.51, Z = 15.36, p < .001 \). As there was significant heterogeneity across assessments, \( Q(23) = 300.99, p < .001, I^2 = 92.35 \), we calculated risk ratios separately for each grade level with subgroup analysis (see Table 3). There was a statistically significant difference in sex ratios between grades, \( Q(2) = 61.74, p < .001 \). There was a moderate overrepresentation of high-achieving males in Grades 4 and 8, but sex ratios increased considerably by Grade 12 to a ratio of 2.13 males to every female student. Although these ratios are still smaller than reported from earlier decades (e.g., Benbow, 1988; Hedges & Nowell, 1995), they remain important targets for educational intervention to encourage and foster high achievement.

Additionally, we tested whether there was a decline in the gender gap for high achievers over time, finding a significant interaction between grade and year of assessment (\( p < .05 \)). To investigate, we performed a meta-regression on year of assessment for each grade level. Although there was a tendency toward slightly smaller sex ratios for Grade 4 students over time, \( Z = -4.45, b = -.0247, CI_{95\%} = [-.0355, -.0138], p < .001 \), there was no association between year of assessment and high achievers in Grades 8 (\( Z = -.37, p = .711 \)) and 12 (\( Z = -1.15, p = .249 \)), indicating stability across the time period examined.

NAEP Assessment of Science

National performance data in science was examined for the period 1996–2011 (see the Appendix for schedule). Overall, the sex difference between males and females was small and comparable to sex differences in mathematics (\( d = .11, Z = 9.15, p < .001 \)). However there was considerable heterogeneity across assessments, \( Q(11) = 328.22, p < .001, I^2 = 96.33 \) (see Figure 2). In order to better explain variability across assessments, we tested whether grade level and year of assessment were potential moderators.

Grade level as a moderator. Using subgroup analysis we partitioned effect sizes across the three grade levels, reducing heterogeneity somewhat. Table 4 presents sex differences in science achievement across each grade level and shows significant differences favoring males across all grades. Although the observed effect sizes were small in magnitude, values for Grade 8 and Grade 12 exceed Hyde’s (2005) criteria for negligible sex differences (\( d = .12 \) and .13, respectively).

Year of assessment as a moderator. Next we performed a meta-regression analysis to test the effect of assessment year as a potential moderator. Contrary to our hypothesis of a declining gender gap in science over time, there was no significant effect of the year of assessment on the magnitude of sex differences in science, \( b = .00, CI_{95\%} = [-.0039, .0057], Z = .37, p = .711 \); nor was there an interaction between year and grade.

Variance ratios. Consistent with previous research, the variability of boys’ performance in science was larger than that of girls’ (see Table 5). Variance ratios across all grades exceeded Feingold’s (1994) criterion for greater male variability and were comparable to that found for mathematics. These variance ratios were also stable across the time period examined, with no association with year of assessment or interaction with grade (\( p > .05 \)).

Gender gaps in high achievers for science. The influence of greater male variability is most readily apparent when looking at sex ratios for attainment of an advanced proficiency standard in science. We calculated the risk ratio of males:females attaining the advanced proficiency standard for science, \( RR = 1.85, Z = 12.81, p < .001 \). As there was significant heterogeneity across assessments, \( Q(12) = 83.32, p < .001, I^2 = 85.63 \), we calculated risk ratios separately for each grade level using subgroup analysis (see Table 5). This reduced heterogeneity somewhat. Sex ratios for students were modest in Grade 4 (1.56) but grew wider for older students in Grade 8 (1.88) and Grade 12 (2.28). There was also a significant difference in science gender gaps between grades, with between-groups heterogeneity, \( Q(2) = 9.05, p = .011 \).

Table 2

Sex Differences in NAEP Mathematics Achievement for Grades 4, 8, and 12

<table>
<thead>
<tr>
<th>Grade</th>
<th>( k )</th>
<th>Cohen’s ( d )</th>
<th>95% confidence interval</th>
<th>Test of null (2-tail)</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9</td>
<td>.07</td>
<td>.06 .09</td>
<td>10.67 &lt;.001</td>
<td>( Q(8) = 90.37, p &lt; .001, I^2 = 91.15 )</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>.04</td>
<td>.03 .06</td>
<td>6.54 &lt;.001</td>
<td>( Q(8) = 28.52, p &lt; .001, I^2 = 71.95 )</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>.10</td>
<td>.08 .12</td>
<td>10.08 &lt;.001</td>
<td>( Q(5) = 10.71, p = ns )</td>
</tr>
</tbody>
</table>

Note: \( k \) denotes the number of assessments conducted for each grade. Effect sizes that exceed Hyde’s (2005) criterion for nontrivial differences (\( d = .10 \)) are highlighted in bold. NAEP = National Assessment of Educational Progress; \( ns \) = nonsignificant.

Table 3

Sex Differences in Variability and Sex Ratios Attaining Advanced Proficiency in Mathematics

<table>
<thead>
<tr>
<th>Grade</th>
<th>Variance ratio</th>
<th>Risk ratio</th>
<th>95% confidence interval</th>
<th>Test of null (2-tail)</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.12</td>
<td>1.51</td>
<td>1.42 1.60</td>
<td>13.71 &lt;.001</td>
<td>( Q(8) = 94.30, p &lt; .001, I^2 = 91.51 )</td>
</tr>
<tr>
<td>8</td>
<td>1.12</td>
<td>1.30</td>
<td>1.23 1.37</td>
<td>9.27 &lt;.001</td>
<td>( Q(8) = 24.71, p = .002, I^2 = 67.63 )</td>
</tr>
<tr>
<td>12</td>
<td>1.15</td>
<td>2.13</td>
<td>1.90 2.38</td>
<td>13.28 &lt;.001</td>
<td>( Q(5) = 8.77, p = ns )</td>
</tr>
</tbody>
</table>
Additionally, we tested whether there was a decline in the gender gap for high achievers over time or an interaction between grade and year. Although there was no significant association with year of assessment overall (Z = 0.84, p = .401), the interaction was significant (p < .05), and we examined effects of year for each level of grade. There was no significant association with year of assessments for Grades 4 (Z = -.13, p = .899) and 12 (Z = -.58, p = .557), but there was a significant trend toward slightly larger science sex ratios in more recent years for students in Grade 8, Z = 2.98, b = -.0260, CI[.95%] [-.0090, .0431], p = .003.

Science domains. Overall science achievement shows only part of the picture, however. NAEP assesses science literacy across three subject domains: physical sciences, earth sciences, and life sciences (see Table 6). If group differences were present across all three domains, sex differences in overall science literacy might be an appropriate target for intervention. However, this was not the case. Although small sex differences were found in physical sciences (d = .13) and earth sciences (d = .17), there were no significant differences for life sciences. The absence of a statistically significant sex difference in life sciences is consistent with the findings of the National Educational Longitudinal Study (Burkam, Lee, & Smerdon, 1997) and the Trends in Mathematics and Science Study (Neuschmidt, Barth, & Hastedt, 2008), which report finding no sex differences in the field of life sciences. We note however that greater male variability was present for all content areas and grades.

There was also considerable heterogeneity of effect sizes across assessments, which may be due in part to the reduced coverage of assessments conducted for science, as well as the smaller sample sizes employed (particularly for Grade 12). Accordingly, moderator analysis was also performed for each science content domain to determine if grade and year effects were present. There was no effect of year of assessment across all three measures or interactions between grade and year of assessment. Although there were no significant effects of grade level for earth and life sciences, there was a tendency for larger sex differences in physical sciences for older students.

### Discussion

Our aim in this study was to evaluate the evidence for sex differences in mathematics and science achievement over a broad span of years and to determine whether these were diminishing over time in response to educational advancements and cultural changes in the roles of men and women (Auster & Ohm, 2000; Wood & Eagly, 2012). The NAEP data set provided an extremely large nationally representative sample of students collected over a wide time span, and it affords a more accurate and reliable test of sex differences in STEM achievement than can be obtained from a single sample. In doing so it extends coverage of the earlier analysis by Hedges and Nowell (1995) to include the most recently available data (1990–2011).

### Sex Differences in Means

In contrast to the analysis by Hyde et al. (2008), which found no difference in a 10-state subset of the national assessment, analysis of the complete NAEP data set found a small but nontrivial mean difference in mathematics favoring males for students in their final year of year of schooling. Furthermore, we extended the analysis to include national testing of science achievement with similar findings. These findings make the claim that sex differences in quantitative reasoning have been eliminated in modern samples somewhat premature, but neither is there evidence of a wide disparity between the performance of the average male and female student. It is also consistent with U.S. performance in international tests of science and mathematics, which have found only small sex differences (Else-Quest, Hyde, & Linn, 2010; Guiso, Monte, Sapienza, & Zingales, 2008; Reilly, 2012).

It is unclear exactly why the earlier meta-analysis by Hyde et al. (2008) on a small subset of testing data found no difference in

---

**Table 4**

**Sex Differences in NAEP Science Achievement for Grades 4, 8, and 12**

<table>
<thead>
<tr>
<th>Grade</th>
<th>k</th>
<th>Cohen’s d</th>
<th>95% confidence interval</th>
<th>Test of null (2-tail)</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower limit</td>
<td>Upper limit</td>
<td>Z</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>.08</td>
<td>.04</td>
<td>.12</td>
<td>3.64</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>.12</td>
<td>.08</td>
<td>.16</td>
<td>6.39</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>.13</td>
<td>.09</td>
<td>.18</td>
<td>6.05</td>
</tr>
</tbody>
</table>

*Note.* Effect sizes that exceed Hyde’s (2005) criterion for nontrivial differences (d ≥ .10) are highlighted in bold. NAEP = National Assessment of Educational Progress.
NAEP mathematics performance, although sex differences in the national data set were somewhat larger. It may be due to educational factors (inherent differences from state to state), from the inclusion of private and public institutions in the national data set, or that when a more representative sample and less selective sample is collected greater sex differences emerge (Hyde et al., 1990). We also note that the magnitude of these mean sex differences in the NAEP was smaller than similar assessments collected in the decades prior to 1990 for mathematics and science (Hedges & Nowell, 1995), which would be consistent with changes predicted by sociocultural perspectives. However, there was no association between the magnitude of the sex difference observed in each assessment and the assessment year, indicating that there was stability across the period of time investigated (1990–2011). That no further change occurred over this time frame would be compatible with biological and psychobiological perspectives. Stability is also consistent with the findings of McGraw et al., (2006), who found no change across a shorter time frame for NAEP mathematics performance. We found meaningful sex differences, but this does not necessarily preclude Hyde’s gender similarities hypothesis as it posits that sex differences in cognitive ability are only small in magnitude.

The data also indicated that there was a developmental trend across both types of quantitative reasoning skills, with smaller effect sizes in elementary school and larger effect sizes in older students. Sex differences in mathematics exceed Hyde’s criterion in Grade 12, whereas sex differences in science achievement reach a nontrivial size in Grades 8 and 12. A prior meta-analysis (Hyde et al., 1990) also found larger sex differences are observed when complex problem-solving tasks are measured, and the mathematics assessment framework increases in complexity during Grades 8 and 12. This is also consistent with developmental literature reporting a widening of the gender gap in quantitative reasoning at around puberty and middle school (Fan, Chen, & Matsumoto, 1997; Hyde et al., 1990; Robinson & Lubinski, 2011), when the saliency of gender roles becomes more prominent as suggested by sociocultural perspectives on gender (Nash, 1979; Ruble, Martin, & Berenbaum, 2006). During adolescence and into early adulthood, gender stereotyping about the sex typing of activities and interests increases at both the explicit and implicit level (Halpern & Tan, 2001; Nosek et al., 2009; Steffens & Jelenec, 2011), which has implications for sex differences in achievement motivation and self-efficacy for mathematics and science (Priess & Hyde, 2010; Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). However, it also coincides with a time of increased hormonal changes as outlined by biological theories (Kimura, 2000), and offering more than speculation as to the origins of sex differences at these developmental periods is therefore difficult.

Of particular interest in our analysis is the observation that mean sex differences were present for some, but not all, of the scientific domains assessed by the NAEP. Despite the considerable sample size there was no sex difference found for biology and life sciences, whereas males and females show equivalent performance (Neuschmidt et al., 2008). Reviews of the literature find that males

### Table 5
**Sex Differences in Variability, and Sex Ratios Attaining Advanced Proficiency in Science**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Variance ratio</th>
<th>Risk ratio</th>
<th>95% confidence interval</th>
<th>Test of null (2-tail)</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower limit</td>
<td>Upper limit</td>
<td>Z</td>
</tr>
<tr>
<td>4</td>
<td>1.09</td>
<td>1.56</td>
<td>1.33</td>
<td>1.83</td>
<td>5.45</td>
</tr>
<tr>
<td>8</td>
<td>1.12</td>
<td>1.88</td>
<td>1.64</td>
<td>2.16</td>
<td>8.95</td>
</tr>
<tr>
<td>12</td>
<td>1.14</td>
<td>2.28</td>
<td>1.88</td>
<td>2.76</td>
<td>8.41</td>
</tr>
</tbody>
</table>

Note. Effect sizes that exceed Hyde’s (2005) criterion for nontrivial differences ($d \geq .10$) are highlighted in bold. NAEP = National Assessment of Educational Progress.
have greater overall interest in science than females do and rate
their aptitude more highly (Osborne, Simon, & Collins, 2003; Weinburg, 1995). But when inquiries are made regarding interest
in specific scientific domains, biology and life sciences show no
significant difference between males and females (Miller, Blessing, & Schwartz, 2006). Rather than indicating any inherent lack of
ability, sex differences in certain but not all domains of science
may reflect different patterns of interest and motivation toward
people-oriented fields (Su et al., 2009), or that other domains are
seen as being less relevant to future career paths (Jones, Howe, &
Rua, 2000; Miller et al., 2006). Alternately, the mathematical
requirements of biology and life sciences may be lower than for
the physical sciences, or there may be reduced sex-typing stereo-
types for this field of study.

High Achievers

Sex difference research often focuses on the performance of the
average student, but considerably less attention is given to sex
differences in the prevalence of high achievers and those factors
that contribute to their success (Wai, Putallaz, & Makel, 2012).
Although only small mean differences in mathematics and science
achievement were found, consistent with prior research the per-
formance of males showed consistently greater variability than that
of female students (Hedges & Nowell, 1995). Greater male vari-
ability in performance is often associated with essentialist biolog-
ic theories of sex differences (Feingold, 1992), but it is also
predicted by differential social and learning experiences afforded
to boys and girls as argued in sociocultural theories of gender. The
combined effect of small mean differences and greater male vari-
ability is then reflected in the sex ratios of students attaining the
high proficiency standard of the NAEP in math and science.
Although there are no established guidelines as to how to interpret
the magnitude of sex ratios, we would suggest that a sex ratio of
over 2:1 (i.e., over twice as many males as females reaching these
standards) should be considered meaningful and nontrivial. Find-
ing a large sex difference in high achievers for mathematics and
science may not be in keeping with a strict interpretation of Hyde’s
(2005) gender similarities hypothesis, but it should be noted that
the hypothesis as it was originally articulated considered only
mean sex differences (Hyde, 2005) and did not speak to gender
imbalances in high achievers. Additionally, there was no overall
effect of year of assessment on tail ratios, though there was a slight
tendency for change in Grade 4 mathematics and Grade 8 science.
It may be the case that changes predicted by sociocultural perspec-
tives operate over a longer time frame or that greater male vari-
ability remains unchanged, as might be predicted by psychobi-
ological theories.

Implications

Although mean sex differences in mathematics and science were
small in magnitude, even small differences in ability level may be
consequential if experienced over time (Eagly, Wood, & Diekmann,
2000; Prentice & Miller, 1992; Rosenthal, 1986). In particular,
they may serve to undermine self-efficacy and interest in tradition-
ally sex-typed subjects such as mathematics and science (Eccles, 2013; Else-Quest, Mineo, & Higgins, 2013). However, this is of less concern than the combined effect of small mean
differences and greater male variability, which leads to large
gender gaps in high achievers for mathematics and science.

Further efforts may be warranted to encourage and cultivate
girls’ interest and aptitude in these subject areas—particularly with
students who have yet to realize their full potential. Many students
have a stereotypically masculine image of mathematics and sci-
ence (Nosek, Banaji, & Greenwald, 2002; Smeding, 2012), and
countering deeply ingrained sex stereotypes is not easily achieved
(Shapiro & Williams, 2012). Although all students receive instruc-
tion in these areas through the school curriculum, parents can
facilitate development of mathematics and science interest and
aptitude by providing early enrichment activities and science
learning experiences equally for daughters and sons (Newcombe &
Frick, 2010). Boys report having more extracurricular experiences
with toys and games that promote science learning (Jones et al.,
2000), and examination of parent-child interactions shows that
parents explain scientific concepts to boys more frequently than to
girls (Crowley, Callanan, Tenenbaum, & Allen, 2001; Diamond,
1994; Tenenbaum & Leaper, 2003). Parents also estimate the
intelligence of sons as being higher than that of daughters, includ-
ing their mathematics intelligence (Furnham, Reeves, & Budhani,
2002), and parental expectations can profoundly impact the self-
efficacy of children (Eccles, Jacobs, & Harold, 1990). Encourag-
ing and supporting daughters who show interest or aptitude in
science to develop their potential may be critical for addressing
gender gaps in high achievers.

The educational environment in which mathematics and science
are taught at school can also have a profound impact on student
learning outcomes (Gunderson, Ramirez, Levine, & Beilock,
2012). Teachers have different beliefs about male and female
students in mathematics, have more frequent interactions with
male than with female students, and have higher expectations in
this field for boys (Li, 1999). Similar findings have been reported
for science education, such as calling more frequently on male
students to answer questions or provide a demonstration (Jones &
Wheatley, 1990). Differential learning experiences for boys and
girls in the classroom are often subtle (Beaman, Wheldall, &
Kemp, 2006) but may be contributing to the development of lower
self-efficacy and less interest in STEM for girls (for a review, see
Gunderson et al., 2012). Individual differences in endorsement of
sex stereotypes about STEM can seriously undermine girls’
achievement in these fields later in life (Schmader, Johns, &
Barquissau, 2004), so it is important that educators send a positive
message about the applicability of mathematics and science skills
to both genders.

A growing body of research also suggests that visuospatial skills
play an important role in the development of quantitative reason-
ing (Nuttall, Casey, & Pezaris, 2005) and that sex differences in
spatial ability may be a mediator (Wai et al., 2009). However, even
brief educational interventions can show marked improvements in
the development of spatial ability in both genders (Uttal et al.,
2013), with evidence of transfer to other quantitative tasks. Many
researchers have advocated for the inclusion of spatial learning
within the school curriculum (Newcombe & Frick, 2010; Priest &
Hyde, 2010), as this would provide benefits to all students and lay
down a solid foundation for the later development of quantitative
reasoning. Contrary to our hypothesis, mean sex differences and
sex ratios of high achievers did not show a decline over the time
period analyzed. Despite societal changes in the roles of men and
women (Auster & Ohm, 2000), this has not translated into diminishing sex differences over time as predicted by social and psychobiological perspectives. The present findings of stable sex differences give further weight to arguments that educational interventions are still required in the interest of gender equity.

**Strengths and Limitations**

The issue of sex differences in quantitative reasoning has been contentious in recent decades, with some researchers arguing that there are considerable differences and others arguing that there are none. By employing a large nationally representative sample such as the NAEP, we can be more confident that the observed sex differences reflect the diversity of socioeconomic status and ethnicity found in the United States, as well as the different educational environments of each state. The statistical technique of meta-analysis makes it possible to aggregate findings from multiple waves of assessment, ensuring that the conclusion reached is not idiosyncratic to a particular assessment year and student cohort. As such it gives greater confidence in estimating the magnitude of sex differences in mathematics and science in U.S. students under the NAEP.

It has also offered the opportunity to test whether the magnitude of said differences is declining and to establish that—at least for the time period analyzed—these are stable across time. It also draws attention to the role that greater male variability can play and the critical importance of examining tail ratios of high-achieving students for a complete test of the gender similarities hypothesis.

Although adding to the existing literature on sex differences, this study is not without limitations. First, it does not provide any information on the causal factors that explain why sex differences emerge. Although researchers have identified a number of biological, psychological, and social factors that contribute to sex differences in quantitative reasoning (Halpern et al., 2007), many researchers agree that a variety of factors are ultimately responsible and advocate a biopsychosocial model of sex differences (Halpern, 2004; Halpern & Tan, 2001). Thus, the findings of a meta-analysis can shed no light on why sex differences emerge and can only document their existence.

Second, our study does not consider other factors, such as socioeconomic background and ethnicity. There is some evidence to show interactions between sex differences and ethnic backgrounds. For example, although sex differences are consistently found for Caucasian and Hispanic students, some studies have failed to find differences for African American samples (Fan et al., 1997; McGraw et al., 2006). Likewise, some studies have found interactions between socioeconomic status and sex differences in early spatial development (Levine et al., 2005), which provides a foundation for quantitative reasoning. Teasing apart such theoretical contributions would be a useful addition to the literature.

Finally, our analysis is limited by the test content being assessed by the NAEP. Previous studies (e.g., Hyde et al., 1990) have noted larger sex differences are found in complex problem solving, but the NAEP includes test items across a range of difficulty levels. International assessments of student ability, such as the Programme for International Student Assessment (PISA), include more challenging test content and find somewhat larger sex differences in mathematics and science for U.S. students than found under the NAEP (Guise et al., 2008; Reilly, 2012). Although these parallel lines of evidence provide a replication of sex differences, they do suggest that the NAEP may underestimate the true effect size of such differences somewhat.

**Summary**

In the present study, we report a meta-analysis of sex differences in mathematics and science achievement in the NAEP, a nationally representative sample of students drawn from public and private institutions from across all states in the United States. Small mean sex differences favoring males were observed in science and mathematics performance, making claims of their absence premature. Further examination of male and female performance across the three domains of science found that males and females were equivalent in performance for life sciences but not for earth and physical sciences. Contrary to our hypothesis, sex differences were not moderated by the year in which students were tested, indicating stability across time. Additionally we found that the performance of males was more variable than that of females, which has implications for the proportion of males to females in the upper-right tail of the ability distribution. Greater male variability may contribute to the disparity in educational outcomes in STEM-related fields, with males being overrepresented in attainment of an advanced proficiency in mathematics and science by a ratio of over 2:1. Further research into the psychological and social factors underpinning these gender gaps is required, as well as educational interventions and support services to help girls realize their full potential in mathematics and science achievement. Counteracting the tendency for initially small sex differences in achievement to be translated into larger sex differences in career choices is likely to require concerted and sustained efforts at many levels.

**References**


Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible
This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

14

REILLY, NEUMANN, AND ANDREWS

me ⫽ female, therefore math not ⫽ me. Journal of Personality and
.1.44
Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala,
stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences, USA, 106,
10593–10597. http://dx.doi.org/10.1073/pnas.0809921106
academic achievement from 1960 to 1994: An analysis of differences in
mean, variance, and extreme scores. Sex Roles, 39, 21– 43. http://dx.doi
.org/10.1023/A:1018873615316
mediator of gender differences on mathematics tests: A biologicalenvironmental framework. In A. M. Gallagher & J. C. Kaufman (Eds.),
review of the literature and its implications. International Journal of
0950069032000032199
President’s Council of Advisors on Science and Technology. (2010).
Prepare and inspire: K-12 science, technology, engineering, and math
Priess, H. A., & Hyde, J. S. (2010). Gender and academic abilities and
preferences. In J. C. Chrisler & D. R. McCreary (Eds.), Handbook of
gender research in psychology (pp. 297–316). New York, NY: Springer.
& Breedlove, S. M. (2010). Salivary testosterone does not predict mental
rotation performance in men or women. Hormones and Behavior, 58,
282–289. http://dx.doi.org/10.1016/j.yhbeh.2010.03.005
Center for Education Statistics.
ONE, 7(7), e39904. http://dx.doi.org/10.1371/journal.pone.0039904
.org/10.1007/s11199-013-0269-0
achievement gaps in mathematics and reading during elementary and
.doi.org/10.3102/0002831210372249
http://dx.doi.org/10.1111/j.1540-4560.1986.tb00247.x
gender differences: The role of stereotype endorsement in women’s
experience in the math domain. Sex Roles, 50, 835– 850. http://dx.doi
.org/10.1023/B:SERS.0000029101.74557.a0
The socialization of sex-differentiated skills and academic performance:
10.1007/BF00289251

Shapiro, J., & Williams, A. M. (2012). The role of stereotype threats in
undermining girls’ and women’s performance and interest in STEM
Smeding, A. (2012). Women in science, technology, engineering, and
mathematics (STEM): An investigation of their implicit gender stereotypes and stereotypes’ connectedness to math performance. Sex Roles,
67, 617– 629. http://dx.doi.org/10.1007/s11199-012-0209-4
Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics
http://dx.doi.org/10.1037/0003-066X.60.9.950
and femininity. In T. B. Sonderegger (Ed.), Nebraska Symposium on
Motivation: Psychology and gender (pp. 59 –95). Lincoln: University of
Nebraska Press.
regarding math and language: Implicit ability stereotypes are selfserving for boys and men, but not for girls and women. Sex Roles, 64,
324 –335. http://dx.doi.org/10.1007/s11199-010-9924-x
Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and
reading achievement are inversely related: Within- and across-nation
dx.doi.org/10.1371/journal.pone.0057988
abilities: Experimental design issues and empirical results. Learning and
Individual Differences, 7, 275–287. http://dx.doi.org/10.1016/10416080(95)90002-0
Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and
people: A meta-analysis of sex differences in interests. Psychological
analyses be undertaken and interpreted? Statistics in Medicine, 21,
http://dx.doi.org/10.1111/j.1745-6924.2009.01169.x
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren,
C., & Newcombe, N. S. (2013). The malleability of spatial skills: A
http://dx.doi.org/10.1037/a0028446
10.1037/0033-2909.117.2.250
domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, 101,
817– 835. http://dx.doi.org/10.1037/a0016127
Are there sex differences, and are the smart getting smarter? Current
Directions in Psychological Science, 21, 382–390. http://dx.doi.org/
10.1177/0963721412455052
Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but
more choice: Individual and gender differences in choice of careers in




Appendix

Table A1
Descriptive Statistics, Effect Sizes, and Variance Ratios for NAEP Mathematics

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
<th>Male M</th>
<th>Male SD</th>
<th>Female M</th>
<th>Female SD</th>
<th>Sample size</th>
<th>Variance ratio</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>4</td>
<td>241.41825</td>
<td>29.76624</td>
<td>239.92438</td>
<td>28.11948</td>
<td>209,000</td>
<td>1.12</td>
<td>0.05</td>
</tr>
<tr>
<td>2009</td>
<td>4</td>
<td>240.61765</td>
<td>29.50272</td>
<td>238.69442</td>
<td>27.86402</td>
<td>168,800</td>
<td>1.12</td>
<td>0.07</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>240.79044</td>
<td>29.43191</td>
<td>238.62343</td>
<td>27.74161</td>
<td>197,700</td>
<td>1.13</td>
<td>0.08</td>
</tr>
<tr>
<td>2005</td>
<td>4</td>
<td>239.11030</td>
<td>28.92446</td>
<td>236.59788</td>
<td>27.81468</td>
<td>172,000</td>
<td>1.08</td>
<td>0.09</td>
</tr>
<tr>
<td>2003</td>
<td>4</td>
<td>236.37463</td>
<td>29.06977</td>
<td>233.41351</td>
<td>27.58066</td>
<td>190,000</td>
<td>1.11</td>
<td>0.10</td>
</tr>
<tr>
<td>2000</td>
<td>4</td>
<td>226.82131</td>
<td>32.34153</td>
<td>224.30827</td>
<td>30.05055</td>
<td>13,800</td>
<td>1.16</td>
<td>0.08</td>
</tr>
<tr>
<td>1996</td>
<td>4</td>
<td>223.73966</td>
<td>31.70661</td>
<td>223.27141</td>
<td>29.95813</td>
<td>6,600</td>
<td>1.12</td>
<td>0.07</td>
</tr>
<tr>
<td>1992</td>
<td>4</td>
<td>220.89259</td>
<td>32.52064</td>
<td>218.52010</td>
<td>30.80918</td>
<td>8,700</td>
<td>1.11</td>
<td>0.07</td>
</tr>
<tr>
<td>1990</td>
<td>4</td>
<td>213.54463</td>
<td>32.73525</td>
<td>212.54085</td>
<td>30.70411</td>
<td>8,900</td>
<td>1.14</td>
<td>0.03</td>
</tr>
<tr>
<td>2011</td>
<td>8</td>
<td>284.45084</td>
<td>37.21046</td>
<td>283.23397</td>
<td>35.12125</td>
<td>175,200</td>
<td>1.12</td>
<td>0.03</td>
</tr>
<tr>
<td>2009</td>
<td>8</td>
<td>283.94915</td>
<td>37.22430</td>
<td>281.85728</td>
<td>35.48711</td>
<td>161,700</td>
<td>1.10</td>
<td>0.06</td>
</tr>
<tr>
<td>2007</td>
<td>8</td>
<td>282.40116</td>
<td>37.40132</td>
<td>280.27550</td>
<td>34.62987</td>
<td>153,000</td>
<td>1.17</td>
<td>0.06</td>
</tr>
<tr>
<td>2005</td>
<td>8</td>
<td>279.61146</td>
<td>37.14541</td>
<td>278.01277</td>
<td>35.43316</td>
<td>162,000</td>
<td>1.10</td>
<td>0.04</td>
</tr>
<tr>
<td>2003</td>
<td>8</td>
<td>278.48139</td>
<td>37.18516</td>
<td>276.63517</td>
<td>35.21715</td>
<td>153,000</td>
<td>1.11</td>
<td>0.05</td>
</tr>
<tr>
<td>2000</td>
<td>8</td>
<td>273.91265</td>
<td>39.25296</td>
<td>272.27437</td>
<td>36.79607</td>
<td>15,000</td>
<td>1.14</td>
<td>0.04</td>
</tr>
<tr>
<td>1996</td>
<td>8</td>
<td>271.43222</td>
<td>38.25208</td>
<td>269.44691</td>
<td>36.62322</td>
<td>7,100</td>
<td>1.09</td>
<td>0.05</td>
</tr>
<tr>
<td>1992</td>
<td>8</td>
<td>268.09776</td>
<td>36.78734</td>
<td>268.70292</td>
<td>35.68133</td>
<td>9,400</td>
<td>1.06</td>
<td>–0.02</td>
</tr>
<tr>
<td>1990</td>
<td>8</td>
<td>263.20994</td>
<td>37.23174</td>
<td>261.87034</td>
<td>34.70190</td>
<td>8,900</td>
<td>1.15</td>
<td>0.04</td>
</tr>
<tr>
<td>2009</td>
<td>12</td>
<td>154.94494</td>
<td>34.89788</td>
<td>151.66908</td>
<td>32.47539</td>
<td>51,700</td>
<td>1.15</td>
<td>0.10</td>
</tr>
<tr>
<td>2005</td>
<td>12</td>
<td>151.31353</td>
<td>35.54736</td>
<td>148.78616</td>
<td>32.35334</td>
<td>15,100</td>
<td>1.21</td>
<td>0.07</td>
</tr>
<tr>
<td>2000</td>
<td>12</td>
<td>301.90598</td>
<td>37.44853</td>
<td>298.52331</td>
<td>33.72126</td>
<td>13,800</td>
<td>1.23</td>
<td>0.09</td>
</tr>
<tr>
<td>1996</td>
<td>12</td>
<td>302.94416</td>
<td>34.96265</td>
<td>300.34237</td>
<td>32.67684</td>
<td>6,900</td>
<td>1.15</td>
<td>0.08</td>
</tr>
<tr>
<td>1992</td>
<td>12</td>
<td>301.33159</td>
<td>34.71171</td>
<td>297.75355</td>
<td>33.04985</td>
<td>8,500</td>
<td>1.10</td>
<td>0.11</td>
</tr>
<tr>
<td>1990</td>
<td>12</td>
<td>297.08056</td>
<td>36.39719</td>
<td>291.48571</td>
<td>34.89335</td>
<td>8,900</td>
<td>1.09</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Note. Effect sizes that are statistically significant at \( p < .05 \) are highlighted in bold. Variance ratios (VRs) above 1.00 indicate greater male variability; VRs below 1.00 reflect greater female variability. NAEP = National Assessment of Educational Progress.

(Appendix continues)
Table A2
Percentage of Male and Female Students Attaining the Advanced Proficiency Level for Mathematics

<table>
<thead>
<tr>
<th>Grade</th>
<th>Year</th>
<th>Male at advanced or higher</th>
<th>Female at advanced or higher</th>
<th>Risk ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2011</td>
<td>7.576799</td>
<td>5.717962</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>6.914833</td>
<td>4.945088</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>6.625340</td>
<td>4.485612</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>5.831723</td>
<td>4.189971</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>4.891417</td>
<td>2.916470</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>3.436696</td>
<td>1.760290</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>3.054985</td>
<td>1.426135</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>2.111726</td>
<td>1.334901</td>
<td>1.58</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>1.685714</td>
<td>0.625990</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grade 4 ratio</td>
<td>1.50</td>
</tr>
<tr>
<td>8</td>
<td>2011</td>
<td>9.216519</td>
<td>7.266763</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>8.801046</td>
<td>7.020203</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>8.103602</td>
<td>5.876708</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>6.731110</td>
<td>5.331733</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>6.118808</td>
<td>4.649840</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>5.917239</td>
<td>4.102644</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>4.296899</td>
<td>3.341992</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>3.172883</td>
<td>2.990450</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>2.366303</td>
<td>1.571801</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grade 8 ratio</td>
<td>1.30</td>
</tr>
<tr>
<td>12</td>
<td>2009</td>
<td>3.520143</td>
<td>1.808225</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>3.062339</td>
<td>1.372165</td>
<td>2.23</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>3.233790</td>
<td>1.363867</td>
<td>2.37</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>2.538683</td>
<td>1.378109</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>2.082941</td>
<td>1.065519</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>2.287939</td>
<td>0.685097</td>
<td>3.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grade 12 ratio</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Table A3
Descriptive Statistics, Effect Sizes, and Variance Ratios for NAEP Science

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
<th>M</th>
<th>SD</th>
<th>Male</th>
<th>M</th>
<th>SD</th>
<th>Female</th>
<th>Sample size</th>
<th>Variance ratio</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>4</td>
<td>150.57607</td>
<td>35.71345</td>
<td>149.40869</td>
<td>34.21378</td>
<td>156,500</td>
<td>1.09</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>4</td>
<td>152.52981</td>
<td>31.78684</td>
<td>148.65937</td>
<td>30.40344</td>
<td>172,500</td>
<td>1.09</td>
<td>0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>4</td>
<td>152.54469</td>
<td>35.28189</td>
<td>147.13105</td>
<td>33.46140</td>
<td>151,000</td>
<td>1.11</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>4</td>
<td>150.85221</td>
<td>33.53692</td>
<td>149.13861</td>
<td>32.39515</td>
<td>173,000</td>
<td>1.07</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>8</td>
<td>154.16130</td>
<td>35.05894</td>
<td>149.21276</td>
<td>33.25938</td>
<td>122,000</td>
<td>1.11</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>8</td>
<td>151.98475</td>
<td>36.14538</td>
<td>147.99001</td>
<td>33.65030</td>
<td>151,100</td>
<td>1.15</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>8</td>
<td>150.48951</td>
<td>36.21289</td>
<td>146.58535</td>
<td>34.31555</td>
<td>173,700</td>
<td>1.09</td>
<td>0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>8</td>
<td>154.35722</td>
<td>36.53652</td>
<td>147.34892</td>
<td>35.03792</td>
<td>15,800</td>
<td>1.09</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>8</td>
<td>150.84548</td>
<td>34.83000</td>
<td>149.12814</td>
<td>32.90619</td>
<td>7,800</td>
<td>1.12</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>12</td>
<td>152.87615</td>
<td>35.79328</td>
<td>147.15407</td>
<td>33.95205</td>
<td>11,100</td>
<td>1.11</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>12</td>
<td>149.01019</td>
<td>34.95658</td>
<td>145.14886</td>
<td>32.58746</td>
<td>22,000</td>
<td>1.15</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>12</td>
<td>147.66712</td>
<td>35.31678</td>
<td>145.17791</td>
<td>32.66081</td>
<td>15,800</td>
<td>1.17</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>12</td>
<td>153.73568</td>
<td>34.34129</td>
<td>147.22612</td>
<td>32.31225</td>
<td>7,500</td>
<td>1.13</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Effect sizes that are statistically significant at \( p < .05 \) are highlighted in bold. Variance ratios (VRs) above 1.00 indicate greater male variability; VRs below 1.00 reflect greater female variability. NAEP = National Assessment of Educational Progress.

(Appendix continues)
<table>
<thead>
<tr>
<th>Grade</th>
<th>Year</th>
<th>Subdomain</th>
<th>Male M</th>
<th>Male SD</th>
<th>Female M</th>
<th>Female SD</th>
<th>Variance ratio</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2009</td>
<td>Earth</td>
<td>151.96979</td>
<td>35.48140</td>
<td>147.94554</td>
<td>34.36776</td>
<td>1.07</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>150.68020</td>
<td>35.76229</td>
<td>149.30413</td>
<td>34.41713</td>
<td>1.10</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>149.02011</td>
<td>35.45167</td>
<td>151.01676</td>
<td>34.48297</td>
<td>1.06</td>
<td>-0.06</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td>Earth</td>
<td>154.62696</td>
<td>34.23227</td>
<td>150.9774</td>
<td>32.83755</td>
<td>1.09</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>152.88549</td>
<td>33.46338</td>
<td>150.37863</td>
<td>31.59997</td>
<td>1.12</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>150.07751</td>
<td>32.25717</td>
<td>147.69223</td>
<td>31.53551</td>
<td>1.05</td>
<td>0.07</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>Earth</td>
<td>155.29923</td>
<td>37.11476</td>
<td>146.93308</td>
<td>35.08561</td>
<td>1.12</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>151.73910</td>
<td>36.73263</td>
<td>149.29376</td>
<td>35.00875</td>
<td>1.10</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>150.59632</td>
<td>36.91456</td>
<td>147.16675</td>
<td>34.57285</td>
<td>1.10</td>
<td>0.10</td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td>Earth</td>
<td>152.68547</td>
<td>35.27448</td>
<td>147.28560</td>
<td>34.50327</td>
<td>1.05</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>150.40359</td>
<td>35.89103</td>
<td>149.59256</td>
<td>34.63000</td>
<td>1.11</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>149.46810</td>
<td>35.33133</td>
<td>150.53822</td>
<td>34.64733</td>
<td>1.04</td>
<td>-0.03</td>
</tr>
<tr>
<td>8</td>
<td>2005</td>
<td>Earth</td>
<td>153.92140</td>
<td>34.84762</td>
<td>147.80144</td>
<td>33.73592</td>
<td>1.07</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>155.57074</td>
<td>35.11631</td>
<td>147.88813</td>
<td>32.92967</td>
<td>1.14</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>153.30020</td>
<td>35.25709</td>
<td>151.58303</td>
<td>33.69989</td>
<td>1.09</td>
<td>0.05</td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td>Earth</td>
<td>152.87875</td>
<td>35.59819</td>
<td>147.08004</td>
<td>34.11659</td>
<td>1.09</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>152.65646</td>
<td>36.30124</td>
<td>147.30709</td>
<td>33.37842</td>
<td>1.18</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>150.69328</td>
<td>35.95506</td>
<td>149.30199</td>
<td>33.95972</td>
<td>1.12</td>
<td>0.04</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td>Earth</td>
<td>152.17593</td>
<td>36.88717</td>
<td>147.55658</td>
<td>34.80591</td>
<td>1.12</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>149.21860</td>
<td>39.26031</td>
<td>142.21619</td>
<td>37.12914</td>
<td>1.12</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>150.17797</td>
<td>36.55163</td>
<td>149.13386</td>
<td>35.07240</td>
<td>1.09</td>
<td>0.03</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>Earth</td>
<td>155.05456</td>
<td>37.39703</td>
<td>148.12953</td>
<td>35.43240</td>
<td>1.11</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>155.15257</td>
<td>38.90284</td>
<td>144.48499</td>
<td>37.61012</td>
<td>1.07</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>153.23782</td>
<td>36.96043</td>
<td>148.93848</td>
<td>35.64924</td>
<td>1.07</td>
<td>0.12</td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td>Earth</td>
<td>151.69663</td>
<td>35.79566</td>
<td>148.25032</td>
<td>34.60768</td>
<td>1.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>151.98523</td>
<td>35.94419</td>
<td>147.95250</td>
<td>33.87181</td>
<td>1.13</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>149.35236</td>
<td>35.93675</td>
<td>150.66815</td>
<td>33.98972</td>
<td>1.12</td>
<td>-0.04</td>
</tr>
<tr>
<td>12</td>
<td>2009</td>
<td>Earth</td>
<td>155.03106</td>
<td>35.30019</td>
<td>145.02064</td>
<td>33.96834</td>
<td>1.08</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>153.45769</td>
<td>36.05441</td>
<td>146.57822</td>
<td>33.58684</td>
<td>1.15</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>150.91495</td>
<td>35.23967</td>
<td>149.09501</td>
<td>34.73195</td>
<td>1.03</td>
<td>0.05</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td>Earth</td>
<td>147.90490</td>
<td>35.52127</td>
<td>142.57167</td>
<td>33.07296</td>
<td>1.15</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>151.21768</td>
<td>37.09497</td>
<td>144.72360</td>
<td>34.38433</td>
<td>1.16</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>147.90856</td>
<td>35.57997</td>
<td>148.15184</td>
<td>33.75475</td>
<td>1.11</td>
<td>-0.01</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>Earth</td>
<td>146.87288</td>
<td>36.40061</td>
<td>142.34830</td>
<td>32.32612</td>
<td>1.20</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>149.10605</td>
<td>37.42376</td>
<td>144.84199</td>
<td>35.29647</td>
<td>1.12</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>147.02286</td>
<td>35.56595</td>
<td>148.34385</td>
<td>32.84964</td>
<td>1.17</td>
<td>-0.04</td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td>Earth</td>
<td>155.85416</td>
<td>35.50721</td>
<td>146.15392</td>
<td>33.06373</td>
<td>1.15</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical</td>
<td>154.14569</td>
<td>36.15524</td>
<td>146.04818</td>
<td>34.04841</td>
<td>1.13</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life</td>
<td>151.20767</td>
<td>34.76459</td>
<td>149.47683</td>
<td>33.20536</td>
<td>1.10</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Note. Effect sizes that are statistically significant at $p < .05$ are highlighted in bold.
Table A5

Percentage of Male and Female Students Attaining the Advanced Proficiency Level for Science

<table>
<thead>
<tr>
<th>Grade</th>
<th>Year</th>
<th>Male at advanced or higher</th>
<th>Female at advanced or higher</th>
<th>Gender ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2009</td>
<td>0.682025</td>
<td>0.512587</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>3.202503</td>
<td>1.853222</td>
<td>1.73</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>4.792086</td>
<td>2.544751</td>
<td>1.88</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>3.418698</td>
<td>2.695513</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grade 4 ratio</td>
<td>1.56</td>
</tr>
<tr>
<td>8</td>
<td>2011</td>
<td>2.203641</td>
<td>0.989545</td>
<td>2.23</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>2.039891</td>
<td>0.967621</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>4.038574</td>
<td>2.349279</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>5.152212</td>
<td>2.799564</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>3.553214</td>
<td>2.530816</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grade 8 ratio</td>
<td>1.88</td>
</tr>
<tr>
<td>12</td>
<td>2009</td>
<td>1.998120</td>
<td>0.827877</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>2.651007</td>
<td>1.266683</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>2.760126</td>
<td>1.390137</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>3.993961</td>
<td>1.346876</td>
<td>2.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grade 12 ratio</td>
<td>2.28</td>
</tr>
</tbody>
</table>
visual-spatial reasoning, and whether sex-role differences would still be present. Thirdly, we employed a verbal fluency language task to examine whether there might an increase in performance (stereotype-lift) in the stereotype priming condition.

The present study was conducted in two phases. In the first, women were assigned to either a masculine or feminine labelling condition to complete the GEFT. It was hypothesized that women in the masculine labelling condition would score lower on the GEFT than those in the feminine labelling condition (Brosnan, 1998). Consistent with the sex-role mediation hypothesis (Nash, 1979; Reilly & Neumann, 2013), it was predicted that women high in masculine sex-role identification (masculine and androgynous groups) would perform better than those low in masculine identification (feminine and undifferentiated). A further question related to the potential interaction of labelling and sex-role identification. It is plausible that the effect of labelling on GEFT performance might be larger for women who identify as highly masculine than for women who identify as less masculine.

In the second phase, women were assigned to either a stereotype-threat inducing condition or a neutral control condition before completing a mental rotation task and a test of verbal fluency. It was hypothesized that women in the stereotype-threat condition would score lower on mental rotation performance than those in the control condition. Consistent with past research (Reilly & Neumann, 2013), we also hypothesized that masculine and androgynous women would score higher on the mental rotation task than feminine and undifferentiated women. Although no \textit{a priori} hypotheses were made, we also sought to determine any interaction (if any) between sex-role category, experimental condition and cognitive ability type. For the verbal fluency language task, it was hypothesized that there would be a stereotype-lift effect, with more words generated in the stereotype-priming condition than in the control condition. Consistent
A frequently observed research finding is that females outperform males on tasks of verbal and language abilities, but there is considerable variability in effect sizes from sample to sample. The gold standard for evaluating gender differences in cognitive ability is to recruit a large, demographically representative sample. We examined 3 decades of U.S. student achievement in reading and writing from the National Assessment of Educational Progress to determine the magnitude of gender differences ($N = 3.9$ million), and whether these were declining over time as claimed by Feingold (1988). Examination of effect sizes found a developmental progression from initially small gender differences in Grade 4 toward larger effects as students progress through schooling. Differences for reading were small-to-medium ($d = -.32$ by Grade 12), and medium-sized for writing ($d = -.55$ by Grade 12) and were stable over the historical time. Additionally, there were pronounced imbalances in gender ratios at the lower left and upper right tails of the ability spectrum. These results are interpreted in the context of Hyde’s (2005) gender similarities hypothesis, which holds that most psychological gender differences are only small or trivial in size. Language and verbal abilities represent one exception to the general rule of gender similarities, and we discuss the educational implications of these findings.

Keywords: gender differences, reading, writing, literacy, sex differences

Supplemental materials: http://dx.doi.org/10.1037/amp0000356.supp

The question of whether males and females differ in cognitive abilities has been the focus of considerable research in recent decades. While there is a general consensus that males and females do not differ in general intelligence (Halpern, 2000), gender differences are commonly observed for more specific cognitive abilities such as visual–spatial ability (Voyer, Voyer, & Bryden, 1995) and language (Miller & Halpern, 2014). However, Hyde (2005) had proposed the gender similarities hypothesis (GSH), which claimed that males and females “are similar on most, but not all, psychological variables. That is, men and women, as well as boys and girls, are more alike than they are different” (p. 581). It holds that most gender differences are small or trivial (close to zero) in magnitude. One exception to this hypothesis may be the gender gap in reading achievement, which is found cross-culturally (Lynn & Mikk, 2009; Reilly, 2012) and exceeds the threshold proposed by Hyde and Grabe (2008, p. 170) for nontrivial gender difference effect sizes ($d \geq .10$). In a recent review, Hyde (2014) remarked that it is “difficult to reconcile” (p. 382) the magnitude of the gender gap observed in reading with other domains of verbal ability (e.g., vocabulary, anagrams), which Hyde and Linn (1988) claimed are typically much smaller.

While the issue of reading is received greater attention, there is a growing body of evidence that males and females also differ in writing ability (Camarata & Woodcock, 2006; Reynolds, Scheiber, Hajovsky, Schwartz, & Kaufman, 2015; Scheiber, Reynolds, Hajovsky, & Kaufman, 2015). Reynolds et al. (2015) noted that the issue of gender differences in writing skills has been overlooked because it is less frequently measured in educational assessments. In cases where writing ability is assessed, researchers should examine gender differences to determine if any meaningful differences occur. Moreover, researchers should compare the size of any differences to those observed with reading assessments when both domains are examined in the same sample.
Some researchers (e.g., Feingold, 1988) have claimed that as a response to societal changes in the status and roles of women, gender differences are declining (see the online supplemental materials for a more detailed discussion of these issues). Gender role attitudes in the United States have changed over time, giving boys the freedom to pursue language-arts fields just as an increasing number of girls now pursue science, technology, engineering, and mathematics fields. Feingold analyzed educational data from 1947 to 1980, showing a decline over time. More recently, Caplan and Caplan (1997, 2016) have questioned whether gender differences in verbal and language abilities even existed at all and were the product of selection bias in samples, while Hyde (2005) has claimed that most gender differences are either small or trivial in size. The current study examines whether historical patterns of gender differences in reading and writing are still present in modern samples and, if so, to determine their magnitude. It presents a meta-analysis of student reading and writing achievement drawn from the National Assessment of Educational Progress (NAEP), a large nationally representative sample of students from the United States conducted by the National Center for Educational Statistics (NCES). Before turning our attention to this dataset, we first present an overview of theoretical perspectives on gender differences in language ability.

**Theoretical Perspectives on Gender Differences in Language Ability**

In their pioneering text *The Psychology of Sex Differences*, Maccoby and Jacklin (1974) presented the first systematic review of the psychological literature on gender differences, arguing that gender differences in verbal ability and language were “well established” (p. 351) and showed a developmental progression toward larger gaps with increasing age. Much of the literature they reviewed focused exclusively on reading ability, rather than considering language proficiency more broadly with higher level tasks such as writing, spelling, and grammar usage. But a number of subsequent studies have also reported gender differences with the largest being spelling and use of grammar (Reilly, Neumann, & Andrews, 2016; Stanley, Benbow, Brody, Dauber, & Lupkowski, 1992).

Theoretical explanations for the emergence of gender differences in reading and language proficiency have been offered. These center around biologically based or sociocultural explanations for gender differences, or combinations of both (Eagly & Wood, 2013; Halpern & Tan, 2001): (a) differential rates of maturation, (b) gender differences in lateralization of brain function, (c) gender differences in variability, (d) gender differences in externalizing behavior and language competence, and (e) gender-stereotyping of reading and language as feminine traits. Each will be discussed in detail next.

**Differential Rates of Maturation**

Girls have a faster rate of maturation and may therefore be attaining greater proficiency than similarly aged boys (Dwyer, 1973), making reading easier and more enjoyable. Such an explanation holds that boys are merely delayed (developmental lag) and boys would attain an equivalent language proficiency given sufficient time. However, this claim is inconsistent with studies showing gender differences in reading that persist into adulthood (Kutner et al., 2007).

**Gender Differences in Lateralization of Brain Function**

Some researchers have claimed that lateralization of brain function for language may differ between males and females (Levy, 1969). It has been claimed that the regions responsible for language tasks are strongly lateralized to the left cerebral hemisphere in right-handed males, but that language regions in females are more likely to be distributed across both the left and right hemisphere (B. A. Shaywitz et al., 1995). Bilateral language function presumably affords some benefits, which could explain the female advantage observed on such tasks. However, empirical support for the Levy hypothesis is mixed (Kaiser, Haller, Schmitz, & Nitsch, 2009), with some neuroimaging studies showing gender differences in lateralization for language tasks (Burman, Bitan, & Booth, 2008; Clements et al., 2006), while others do not (Wallentin, 2009).
Gender Differences in Variability

One explanation for lower reading and language proficiency in males is the greater male variability effect, which states that males show greater variability in cognitive performance across all cultures (Feingold, 1992; Machin & Pekkarinen, 2008). Even if there were no gender differences in group means, the consequence of greater male variability is that males will be overrepresented at the extreme left tail of the ability distribution, which Hawke, Olson, Willcut, Wadsworth, and DeFries (2009) argued explains why gender ratios of poor readers favors females. Boys are also overrepresented in populations with reading impairment, dyslexia, attention disorders, and mental retardation suggesting that there may be a gender-linked neurological contribution (Halpern, Beninger, & Straight, 2011). While explanations for the greater male variability hypothesis in intelligence have been made by evolutionary psychologists (Geary, 2010), few specific evolutionary theories have been proposed for verbal and language abilities (Geary, Winegard, & Winegard, 2014), perhaps because these are more recent in an evolutionary sense.

Gender Differences in Externalizing Behavior and Language Competence

Other researchers have argued that gender differences in externalizing behavior may also partly explain a greater female language competence (Limbrick, Wheldall, & Madelaine, 2011). Clinicians identify more boys than girls with externalizing behavior and attention disorders (McGee, Prior, Williams, Smart, & Sanson, 2002) which have both been associated with reading and language impairment. For example, in a longitudinal sample from the United States, Rabiner and Coie (2000) reported that attention-impairment and externalizing behavior measured in kindergarten predicted later reading impairment in fifth grade. Other studies have followed children over longer time frames. In a longitudinal study of child development in Australia, Smart, Prior, Sanson, and Oberklaid (2001) found that externalizing behavior problems at age 7 predicted the severity of later reading and spelling difficulties at ages 13–14, even after controlling for intelligence and socioeconomic status. Although also present in girls, Smart et al. found that the association between externalizing behavior and reading impairment was significantly stronger in boys. Such an association is not necessarily causal, and may well be reciprocal in nature. Within the context of the educational environment, inattention and behavior problems may result in additional educational setbacks, as such problems can interfere with learning as well as lower academic motivation and rapport between teacher and student. But it is equally plausible that these conditions are related to a common neurobiological factor (Berninger, Nielsen, Abbott, Wijsman, & Raskind, 2008).

Gender-Stereotyping of Reading and Language as Feminine Traits

Kagan (1964) first observed that children readily classify social behaviors and even intellectual tasks as either masculine or feminine in nature, based on shared cultural beliefs about gender roles. Reading and language are generally regarded as feminine in nature (Plante, de la Sablonnière, Aronson, & Théorêt, 2013), and gender stereotypes about language are held by both males and females (Halpern, Straight, & Stephenson, 2011). The process by which a child acquires stereotypically masculine and feminine personality traits is termed sex-typing (Bem, 1981). Highly sex-typed individuals are motivated to keep their behavior and self-concept consistent with traditional gender norms (Martin & Ruble, 2010; Nash, 1979). The rigidity of sex-roles may translate into decreased reading interest and motivation for some boys if there is a perceived incompatibility between reading and masculine norms. Reading motivation is proposed as playing a strong role in later reading achievement, with boys reporting lower reading motivation and interest (Marinak & Gambrell, 2010; Mucherah & Yoder, 2008). Lowered reading motivation is reflected in the amount of leisure time spent on reading (Moffitt & Wartella, 1991), leading to differential levels of practice between boys and girls. Girls in elementary school also report more positive competence beliefs than boys for reading and language tasks (Eccles, Jacobs, & Harold, 1990).
Large-Scale Assessments of Reading and Writing Achievement

One of the difficulties in evaluating research in the field of gender differences in cognitive ability comes from the use of sampling methods, and the potential for selection bias. It is not normally feasible to sample every male and female in a given population. Researchers thus often take a sample group of participants and then use statistics and probability to draw an inference about the underlying population. Hedges and Nowell (1995) note that this approach can be problematic for two reasons. First, as noted earlier, the greater male variability effect results in a greater number of male high and low achievers at the top and bottom of the ability distribution, respectively (Hawke et al., 2009; Machin & Pekkarinen, 2008). Greater variability may present a distorted picture of the underlying population which is magnified in highly selected samples (Becker & Hedges, 1988). Second, demographic factors such as socioeconomic status, ethnicity, and rural versus urban residence can greatly influence cognitive ability (Fernald, Marchman, & Weisleder, 2013; Hanscombe et al., 2012), which may further limit the generalizability of a convenience sample.

For this reason, the gold standard for research is to recruit a large sample that is representative of the population under investigation (Hedges & Nowell, 1995), in terms of gender, ethnicity, socioeconomic status, geographical region, and so forth. This approach increases confidence in the validity of any conclusions made about specific groups, such as males and females. Another reason why selection bias may be problematic in the context of gender differences in reading and writing is that when investigating specific subgroups (such as students that have been identified as poor readers), it is difficult to determine the underlying prevalence of males and females due to the issue of a gendered referral bias. Shaywitz, Shaywitz, Fletcher, and Escobar (1990) noted that more boys than girls are identified as poor readers by educational institutions, but when epidemiological studies investigate reading impairment in the community girls and boys approach an equal representation (Hawke et al., 2009; Jiménez et al., 2011). The implication here is that the prevalence of reading impairment in girls may simply just be underreported, and that there may be a referral bias for boys. In order to test such a claim, a study would need to recruit a large, nationally representative sample and administer a standardized reading assessment.

One such source is NAEP, which is conducted by NCES, part of the U.S. Department of Education. It has the added advantage that new waves of assessment have been conducted over several decades without major changes to the reading and writing frameworks so that temporal trends can be investigated. Before turning our attention to this analysis, we first review previous studies that have recruited nationally representative samples of males and females to investigate gender differences in reading and writing.

Gender Differences in Reading

Hedges and Nowell (1995) reported the largest study of gender differences in achievement scores ever conducted, across a wide range of content areas using nationally representative samples from the United States. These included student assessments of reading proficiency conducted by NAEP reported from 1971 to 1992. They found that girls showed significantly higher scores for tests of reading in each year of assessment, with effect sizes ranging from $d = -.18$ to $-.30$. Furthermore, they found that the performance of boys was more variable than that of girls with an average variance ratio (VR) of 1.12. This variability resulted in an overrepresentation of boys as poor readers. The researchers also examined data from a number of other data sets that recruited nationally representative stratified samples. Across these other data sets, Hedges and Nowell found similarly sized gender differences and greater male variance. They also reported that the ratio of boys to girls in the bottom 10% of reading comprehension (i.e., poor readers) ranged from 1.07 to 1.75, which paralleled that found in the NAEP data. Thus, there were both mean gender differences in reading ability and an overrepresentation of boys who are poor readers.

While pioneering at the time it was published, a serious limitation of Hedges and Nowell’s (1995) analysis was that they only examined NAEP data from students near the end-point of their education, aged 17, and did not investigate whether gender differences were still present in younger students. Developmental differences are an impor-
Gender Differences in Writing

As noted earlier, there are a limited number of studies that have investigated gender differences in writing ability, and the number of studies recruiting representative samples are even fewer. Nowell and Hedges (1998) reported a more detailed analysis of NAEP writing data from the period 1984–1994, finding substantial gender differences in writing (ranging from $d = -0.49$ to $-0.55$), greater male variability, and that gender ratios for students falling in the bottom 10th percentile were between 2.6 and 3.3 males to every female (Nowell & Hedges, 1998, p. 38). At present, there has been no subsequent meta-analysis published investigating gender differences in NAEP writing assessments.

Two other prominent studies have investigated gender differences in writing with large representative samples. Camarata and Woodcock (2006) presented data from the normative samples of the Woodcock-Johnson cognitive and achievement batteries, a large representative sample of males and females aged 5 through to 79. Females scored significantly higher in writing achievement, with an average effect size across the life span of $d = -0.33$. More recently, Scheiber et al. (2015) analyzed a large nationally representative sample of adolescents and young adults completing the Kaufman Test of Educational Achievement—Second Edition Brief Form, which measures participants across reading, writing, and mathematics. While no difference was found in mathematics, females scored higher than males on the tests of reading and writing ability. The effect size for reading was small ($d = -0.18$), but the effect size for writing ($d = -0.40$) was twice as large as that for reading. Given the appreciable gender differences found in these samples, it seems justifiable to expect a similarly sized effect in NAEP data for writing tasks.

The Present Review

We sought to investigate whether the historical patterns of gender differences in reading and writing reported by Hedges and Nowell (1995) would be replicated for children growing up in more recent decades. Consistent with previous research, we hypothesized that gender differences in reading and writing achievement would be present. Based on the claim made by Feingold (1988) and Caplan and Caplan (2016) that gender differences in cognitive ability are decreasing, we also hypothesized that there would be a significant negative association between year of assessment and effect size, such that gender differences would show a decline over time. Given the large sample size employed by NAEP and that data from several decades of testing were available, the analysis would have strong statistical power to detect an effect. Hyde and Grabe (2008) have advocated that a threshold of evidence higher than statistical significance be adopted because although a very large sample size might yield statistically significant differences, the actual size of the effects might be trivial. Therefore we adopted the research practice recommended by Hyde and Grabe (2008, p.
170) and determined a priori that effect sizes smaller than $d = .10$ are characterized as trivial in size, even if they met the threshold for statistical significance. We used Cohen’s (1988) recommendation that effect sizes around $d = .20$ be regarded as small, while around $d = .50$ be medium.

**Method**

**National Assessment of Educational Progress Data Source**

The NAEP is a project of NCES, part of the U.S. Department of Education. The NAEP is used to track student achievement over time in fourth-, eighth- and 12th-grade at the state and national level of the United States. It measures student achievement in reading, mathematics, science and a variety of other subject areas. National and state performances are reported annually in a series of reports titled “The Nation’s Report Card” (see http://nationsreportcard.gov/). This information is of use to parents, educators, and policymakers. However such reports only indicate that gender differences are statistically significant, without providing any context about the size of such differences or gender ratios of poor/advanced readers and writers.

NAEP data is also publically available so that it can be used by researchers to conduct secondary analysis, via the NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepdata/). The sampling frame employed by NAEP is all students in the target grades (Grades 4, 8, and 12) in each of the 50 states of the United States, drawn from both public and private educational institutions. School and student responses are appropriately weighted to draw a nationally representative estimate of the target population that reflects student demographics such as socioeconomic status of school district, ethnicity, rural versus urban location and gender. For inclusiveness, the sampling frame also includes students with disabilities or English language learners. Additional information on the sampling methodology employed is available from the NAEP website (http://nces.ed.gov/nationsreportcard/about/samplesfaq.aspx).

Content for the reading assessment includes reading comprehension of a variety of different passages and genres (including information reports, stories, poetry and essays), as well as an understanding of vocabulary. Content for the writing assessment includes persuasive, informative, and narrative writing in response to stimuli material. Additional information on reading and writing frameworks in each grade level is available from the NAEP website.

**Schedule of Assessment**

Reading and writing assessments are conducted periodically, in adherence with the NAEP schedule. Reading assessments are given greater priority than writing and occur every 2 to 3 years (1988, 1990, 1992, 1994, 1998, 2000, 2002, 2003, 2005, 2007, 2009, 2011, 2013, 2015), with greater coverage given for students in Grades 4 and 8. Writing assessments occur approximately every 4 to 5 years (1998, 2002, 2007, 2011), and usually with a smaller sample size than the reading assessments. We also included archived data from the 1988, 1990, 1992, and 1996 writing assessments so that both dependent variables were assessed across the same time frame. All assessments from 1988 onward were included in the analysis.

**Participants**

National performance data for NAEP Reading assessments were examined from the period 1988–2015, with a combined total sample size of 3.035 million students. Testing data for the NAEP Writing assessments were examined for the period 1988–2011, with a combined total sample size of 934,800. Students provided deemed consent through their participation in each wave of assessment. This study used published archival data and did not recruit participants directly.

**Meta-Analytic Procedure**

Effect size statistics are presented as the mean difference between boys and girls in standardized units, commonly referred to as Cohen’s $d$ (Cohen, 1988). The meta-analysis employed a random effects model. Heterogeneity across samples was indicated by the $I^2$ statistic, representing the percentage of variation across samples attributed to genuine heterogeneity and not chance. We also investigated whether there were developmental differences in the magnitude of the gender gap across the three grade levels using subgroup analysis, and whether the year of testing was a potential moderator using metaregression (Kelley & Kelley, 2012). Full details of the methodology employed in our analysis are reported in the online supplemental materials.

**Results**

**Gender Differences in Reading Achievement**

Girls showed significantly higher reading scores than boys across every wave of assessment and in every grade, with an overall effect size of $d = -.27$, 95% confidence interval (CI) [-.29, -.25], $Z = -26.08$, $p < .001$ (see Figure 1). Gender differences significantly exceeded the predetermined cutoff ($d \leq .10$) advocated by Hyde and Grabe (2008) by a factor of 2.7. There was also significant heterogeneity in effect sizes, $Q(36) = 2594.45$, $p < .001$, $I^2 = 98.61$, indicating considerable variation across assessments. To better explain the variability in effect sizes, we...
investigated whether grade level or year of assessment were potential moderators of the gender difference.

**Grade level.** Table 1 presents comparisons between males and females in reading achievement across the three grade levels assessed by NAEP. There was a statistically significant difference between groups, $Q(2) = 148.49, p < .001$, with a tendency toward larger differences between boys and girls in older students. The initial gender difference in reading achievement was small in Grade 4 ($d = -.19$), but grew larger in Grade 8 ($d = -.30$) and Grade 12 ($d = -.32$).

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Grade</th>
<th>k</th>
<th>Cohen’s $d$</th>
<th>Lower limit</th>
<th>Upper limit</th>
<th>Test of null (two-tail)</th>
<th>VR</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>4</td>
<td>14</td>
<td>-.19</td>
<td>-.21</td>
<td>-.18</td>
<td>-25.00</td>
<td>&lt;.001</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>13</td>
<td>-.30</td>
<td>-.32</td>
<td>-.29</td>
<td>-39.07</td>
<td>&lt;.001</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>10</td>
<td>-.32</td>
<td>-.34</td>
<td>-.30</td>
<td>-32.87</td>
<td>&lt;.001</td>
<td>1.22</td>
</tr>
<tr>
<td>Writing</td>
<td>4</td>
<td>7</td>
<td>-.42</td>
<td>-.47</td>
<td>-.37</td>
<td>-17.55</td>
<td>&lt;.001</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>9</td>
<td>-.62</td>
<td>-.66</td>
<td>-.58</td>
<td>-29.94</td>
<td>&lt;.001</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>9</td>
<td>-.55</td>
<td>-.59</td>
<td>-.51</td>
<td>-26.40</td>
<td>&lt;.001</td>
<td>1.07</td>
</tr>
</tbody>
</table>

**Year of assessment.** Next, we performed a metaregression on reading achievement, using the year of assessment as a predictor. There was no significant effect of year of assessment, $Z = .79, b = .0001, 95\% CI [-.001, .003], p = .425$, which is inconsistent with the hypothesis of a declining gender difference over time.

**Variance ratio.** Consistent with previous research there was greater male variability present in every sample tested, although it sometimes felt just short of Feingold’s threshold (VRs $> 1.1$) for individual years. Mean VR ratios were calculated for each grade. All grades exceeded Feingold’s

![Figure 1](https://example.com/figure1.png)
critical value, with progressively higher variance in older students.

**Gender ratios for poor and gifted readers.** In order to evaluate the combined effect of mean gender differences and greater male variability on poor and gifted readers, we examined the gender ratios of readers falling below the “basic” proficiency standard defined by NAEP as well as those exceeding the “advanced” proficiency standard. The analysis examined the risk ratio of males to females attaining these levels. Equivalent proportions of boys and girls at a particular achievement level would be indicated by a risk ratio of 1.00. Higher risk ratios (i.e., $>1.00$) would indicate an overrepresentation of boys attaining this standard, while lower risk ratios (i.e., $<1.00$) would reflect an overrepresentation of girls at a particular standard.

The weighted risk ratio for poor readers was 1.39, 95% CI [1.34, 1.44], $Z = 19.82, p < .001$. The subgroup analysis is reported in Table 2. As can be seen, more boys than girls were poor readers, which reached a ratio of 1.54 times as many boys as girls falling below the minimum standard of literacy by Grade 12. The effect was reversed for advanced readers, with more girls than boys achieving the advanced literacy standard. Additionally the concentration of males in the lower left tail of the distribution was higher than the concentration of females at the upper right. The weighted risk ratio for advanced readers was 0.55, 95% CI [0.52, 0.59], $Z = -17.28, p < .001$, with subgroups also reported in Table 2. Expressed in a metric that may be more intuitive for nonstatisticians, by the time students reach Grade 12 there are almost twice as many girls than boys that reach the advanced standard of reading proficiency. Moderator analysis showed a slight tendency toward smaller gender gaps in poor readers over time ($Z = -2.55, p = .010$), but larger gender gaps in advanced readers over time ($Z = 2.31, p = .021$).

**Gender Differences in Writing Achievement**

Next we examined the gender difference in writing achievement for the period 1988–2011. Overall, the gender difference between males and females in writing was larger than that found for reading, $d = -.54$, 95% CI $[-.57, -.51]$, $Z = -36.14, p < .001$ (see Figure 2). Gender differences in writing exceeded the predetermined cutoff ($d \leq .10$) advocated by Hyde and Grabe (2008) by a factor of 5.4. There was also significant heterogeneity in effect sizes, $Q(24) = 974.07, p < .001, I^2 = 97.54$, indicating considerable variation across assessments. In order to better explain the variability in effect sizes, we investigated whether grade level or year of assessment were potential moderators of the gender gap. An additional factor introducing heterogeneity may be the changes in writing frameworks (new frameworks were introduced in 1988, 1998, and 2011) and the marked variability in sample sizes for more recent assessments.

**Grade level.** Table 1 presents comparisons between males and females in writing achievement across the three grade levels assessed by NAEP. The difference between grades was statistically significant, $Q(2) = 42.01, p < .001$, with a tendency toward a smaller initial gender difference in writing proficiency for students in Grade 4. The initial gender difference in writing was medium-sized in Grade 4 ($d = -.42$), but grew larger in Grade 8 ($d = -.62$) and Grade 12 ($d = -.55$).

**Year of assessment.** We performed a metaregression on writing achievement, using the year of assessment as the predictor. There was no significant effect of year, $Z = -1.85, b = -.004, 95% CI [-.001, .001], p = .063$, indicating stability in effect sizes across historical time.

**Variance ratio.** In examining the variance ratios presented in Table 1, there was minimal support for greater male variability with all grades falling short of Feingold’s threshold.

**Gender ratios for poor and gifted writers.** In order to evaluate the joint effect of greater male variability and mean gender differences on poor and gifted writers, we examined the gender ratios of readers below the ‘basic’ achievement level. Writing proficiency levels attained were not pub-

---

### Table 2

**Risk Ratio for Poor and Advanced Proficiency Level Readers and Writers, Across Grade Levels**

<table>
<thead>
<tr>
<th></th>
<th>Poor readers</th>
<th></th>
<th>Advanced readers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95% Confidence interval</td>
<td>Test of null</td>
<td>Risk ratio</td>
<td>95% Confidence interval</td>
</tr>
<tr>
<td>Outcome</td>
<td>Grade</td>
<td>Risk ratio</td>
<td>Lower limit</td>
<td>Upper limit</td>
</tr>
<tr>
<td>Reading</td>
<td>4</td>
<td>1.22$^{ab}$</td>
<td>1.20</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1.45$^{b}$</td>
<td>1.43</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1.54$^{bc}$</td>
<td>1.50</td>
<td>1.58</td>
</tr>
<tr>
<td>Writing</td>
<td>4</td>
<td>2.01</td>
<td>1.55</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2.27</td>
<td>1.89</td>
<td>2.72</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>2.21</td>
<td>1.84</td>
<td>2.65</td>
</tr>
</tbody>
</table>

*Note.* Three planned contrasts between grades were conducted, with a Bonferroni correction applied to control family-wise Type I error rate. $^a$ Grade 4 versus 8 significant. $^b$ Grade 4 versus 12 significant. $^c$ Grade 8 versus 12 significant.
lished for the archived reports (1988–1996), and were therefore excluded from analysis.

The weighted risk ratio for poor writers was 2.19, 95% CI [2.00, 2.40], \(Z = 17.06, p < .001\), with subgroup analysis reported for grades in Table 2. As can be seen by the table, there were twice as many boys falling into the category of poor writing than girls. However, the effect is reversed for advanced readers with more girls achieving the advanced standard for written expression, with a weighted risk ratio of 0.30, 95% CI [0.25, 0.36], \(Z = 13.29, p < .001\). In other words, by the time students reach Grade 12 there are over 2.54 times as many girls than boys that attain the advanced standard of writing proficiency. Moderator analysis showed no change in gender ratios for poor writers over time (\(Z = .15, p = .881\)), or advanced (\(Z = 1.80, p = .071\)).

**Discussion**

Annual reporting of NAEP data had noted girls performed significantly higher than boys, but failed to provide estimates of how large these differences were. By calculating an effect size, we can hold evidence of gender differences to a much higher standard than mere statistical significance by examining whether the effect is *practically* significant. While a focus on mean gender differences is important, we also considered its combined effect with greater male variability on the gender ratios at the lower left (poor readers/writers) and upper right (advanced readers/writers) tails of the ability distribution. Both measures (effect size for mean gender differences, gender ratios of low/high achievers) provide a more comprehensive perspective than simply examining probability values. Further, the detailed records kept for NAEP testing data offered a window into the past to examine how boys and girls have fared in reading and writing achievement over historical time (both developmentally across grades, and cohort effects across historical time).

**Reading Proficiency**

Girls significantly outperformed boys in reading ability across all grades, with a tendency toward larger effect sizes in high school than primary school (Grades 8, \(d = -.30\) and 12, \(d = -.32\)). These exceed Hyde’s criterion by a factor of 3, and fall in the small-to-medium effect size category proposed by Cohen (1988). They are also comparable to effect sizes for American students in international assessments such as PISA (Reilly, 2012) and the Progress in International Reading Literacy assessment (PIRLS; Mullis, Martin, Gonzalez, & Kennedy, 2003), where small to medium effect sizes were found. There was also no evidence of a decline in the magnitude of effect sizes over time as had

![Figure 2. Histogram of effect sizes for the difference between boys and girls in writing achievement. All effect sizes fall to the left of the line of no effect and exceed Hyde’s criterion for nontrivial gender differences.](image-url)
been hypothesized, though it is possible that this might be detectable over a longer passage of historical time. However to compare these results to other standardized tests of reading and writing would be problematic, and introduce a methodological confound in test content, level of difficulty, and sampling size.

How ought we interpret the practical impact of such gender differences in reading? Rosenthal and Rubin (1982a) developed the binomial effect size display (BESD) to illustrate the practical impact of such differences for nonstatisticians (such as parents and educators), especially for students falling near the middle of the distribution. This metric shows the percentage of males and females that meet or exceed an average score. Represented in the BESD format, the likelihood of being average or higher in reading ability for a student at the end of high school increases from 42.1% for boys to 57.8% for girls, a not insubstantial amount.

We can also contextualize this by considering the size of gender differences for other types of cognitive ability, such as science, technology, engineering, and mathematics achievement. While considerable research focuses on the gender gap in mathematics and science, the effect sizes for reading are more substantial—over twice the size as that found in comparable NAEP assessments of mathematics (McGraw, Lubienski, & Strutchens, 2006; Reilly, Neumann, & Andrews, 2015). Thus, it is important to acknowledge female strengths as well as those areas where males perform higher. Claims by researchers such as Caplan and Caplan (1997, 2016) that cognitive differences are disappearing are therefore premature, but neither does it support a claim that boys and girls are radically different in reading literacy and would benefit from gender-segregated instruction as is claimed by same-sex advocates.

However, effect size statistics only represent the typically performing girl or boy. When we examined students that fall below the minimum proficiency standard, far more boys than girls fall into this category across all grades and more importantly by the end of high school by a factor of 1.5. This imbalance in a representative sample contradicts the claim made by Shaywitz et al. (1990) that a greater diagnosis of reading impairment in boys is merely the result of a gendered referral bias. A completely different pattern was found for advanced readers though, with far more girls attaining this level of proficiency (by a factor of almost 2).

The pattern of results shows that at all levels of the ability distribution, girls significantly outperform boys in reading achievement. Hyde (2005) had proposed the GSH, arguing that most mean gender differences are small or trivial in magnitude. A limitation of that hypothesis though is that it focuses exclusively on mean gender differences and effect sizes, while ignoring evidence from the upper and lower tails of the ability distribution. Taken together, it would appear there are gender differences in reading favoring girls across all levels of ability distribution, with these being small (and more similar) in the middle of the distribution but much larger (and impactful) at the tails. Furthermore, these gender differences are found in younger students, as well as older ones. We did not find strong support for a greater male variability hypothesis because the larger number of low scoring boys was offset by the higher number of high scoring girls.

Writing Proficiency

As was expected from previous studies (e.g., Hedges & Nowell, 1995; Reynolds et al., 2015), girls significantly outperformed boys in writing ability across all grades and assessment waves. The magnitude of effect sizes was higher than that found for reading, with effect sizes falling into the medium size range by Cohen’s (1988) conventions. Comparisons between boys and girls were slightly smaller in Grade 4 \((d = - .42)\), but this gender difference widened for older students (Grades 8, \(d = -.62\); Grades 12, \(d = -.55\)). Represented in the BESD format, the likelihood of being average or higher in writing ability for a student at the end of high school increases from 36.7% for boys to 63.3% for girls (i.e., a minority of boys attain this standard, but the majority of girls do). Furthermore, when examining the association between effect size and year of assessment, there was no decline in the magnitude of effect sizes as predicted.

At the lower end of the ability distribution, boys were greatly overrepresented by a factor of 2 or more which grew slightly larger for older students. Just as with reading, there was a reversal of gender ratios for students attaining an advanced writing proficiency, with girls greatly overrepresented by a factor of 2 or more. These results are consistent with the position held by Reynolds et al. (2015) who argued that a gender difference in writing is an exception to the GSH and cannot be easily dismissed as a small or trivial difference.

Why might the effect size for writing be larger than the effect size for reading? Writing represents a more challenging task, and larger gender differences are typically found as the complexity of the task increases. While reading is a passive task, writing is a generative task that draws on other components of verbal and language abilities that typically show larger gender differences. For example, it requires careful organization of ideas and the production of material that is clearly expressed, and grammatically accurate. Research shows that females score significantly higher on standardized tests of spelling and of grammar, with medium-sized effects (Stanley et al., 1992). Finding the right words to express a particular concept or nuance is also a demanding task for writers, and draws on verbal fluency (where females also show significantly higher performance than males). Halpern and Tan (2001) noted that effect sizes for verbal fluency fall in the medium to large range. All of these verbal skills can be improved with sufficient practice
and instruction, however, highlighting the importance of these basic skills in a crowded educational curriculum.

**Gender Similarities Hypothesis**

Hyde (2005) has proposed the GSH, which claims that most—but not all—psychological gender differences are small or trivial in size. Zell, Krizan, and Teeter (2015) used the technique of metasynthesis to test this claim, finding that most effect sizes were small. However, they also noted a number of important exceptions (see Zell et al., 2015, Table 3). The gender differences observed in the present study for reading (small-to-medium) and writing (medium-sized) also represent exceptions that may have been overlooked because a meta-analysis on the NAEP dataset had yet to be published for modern samples. The identification of new areas where meaningful gender differences remain does not invalidate the GSH, but does serve as a prompt for further investigation.

**Educational Implications**

What might be the educational implications of such a gender gap in reading and writing proficiency during primary and high school years? While information in the classroom environment is often presented verbally, students are expected to read textbooks and literacy material as independent reading. Difficulties in reading would be a serious impediment, particularly if reading takes boys longer or if they are unable to gain a deep understanding of the text. While it has been known for some time that boys are overrepresented in reading disabilities and formal diagnoses of dyslexia (Berninger et al., 2008; Rutter et al., 2004), this pattern of results suggests a more general reading deficiency for the typical male student. Written communication is also important during the high school years, as this format is commonly adopted in the format of essays or laboratory reports. While boys tend to perform better than girls on standardized tests, girls tend to achieve significantly higher grades during schooling (Duckworth & Seligman, 2006; Voyer & Voyer, 2014), and it is possible that the assessment format (exam vs. written assignment) may be a contributing factor. While the issue of gender differences in reading ability has been the focus of much research, gender differences in writing ability may have been previously underestimated by researchers and educators. The magnitude of the gender gap in writing ability is sufficiently large that it may warrant educational interventions and further research on etiology. It may also be reflective of a more general language deficiency (rather than just an issue with reading), as other studies have also reported pronounced gender differences in grammar and language usage (e.g., Stanley et al., 1992).

While the existence of a gender gap in reading and writing during compulsory schooling is troubling, the educational implications for students considering pursuing tertiary education are potentially compounded. In a review of gender inequalities in education, Buchmann, DiPrete, and McDaniel (2008) note that women enroll in college and universities at a much higher rate than their similarly aged male peers, achieve higher grades on average than males, and have a higher rate of degree completion (Buchmann & DiPrete, 2006). This pattern is mirrored across most OECD countries and is not confined to the United States (OECD, 2016). The transition from secondary to tertiary education can be difficult for many students, because it involves independent learning and considerable hours of study outside of classroom contact time. The ability to read textbooks and assigned readings is a crucial part of learning. Although especially poor readers are less likely to pursue tertiary studies, the gender gap in reading appears to be present in average students as well, though smaller. In addition, the ability to communicate verbally in a written format takes on increasing importance, as producing reports and essays are a common form of student assessment. Systemic gender gaps in writing might leave male students significantly underprepared for tertiary admission, and offer a partial explanation for why females on average achieve higher grades in their tertiary studies (Voyer & Voyer, 2014) and have higher completion rates (Buchmann et al., 2008).

Parents, educators, and policymakers may wonder what to make of gender differences in reading and writing, and what changes might be made to address them in the interests of equality of educational outcomes. It would be a mistake to take evidence from this study to argue that boys and girls learn in fundamentally different ways, require different styles of teaching, or would benefit from same-sex schooling. Scientific literature is clear about the negative effects of highlighting gender in this way (Halpern, Eliot, et al., 2011), and how treating a particular demographic group (i.e., just boys) can serve to undermine their confidence and motivation to improve. While attention has been paid in the past to early intervention for reading, educational interventions for writing may be warranted, and a greater focus on writing tasks in the curriculum to provide additional opportunities to practice writing skills and provide feedback to students. These should be offered broadly to all students—while the findings of this study suggest that boys would benefit from these initiatives, these results also suggest that many girls would similarly benefit.

**Future Directions for Research and Limitations**

While this study documents the existence and magnitude of gender differences in reading and writing, it cannot shed any light on their etiology and which biological and socialization factors most contribute to their development (Eagly & Wood, 2013). Most researchers advocate a biopsychosocial model of gender differences (Halpern, 2000), but some tentative conclusions may be drawn from the generalizability of gender differences in language outcomes across sam-
amples, historical time periods, and cultures. Our study replicates findings reported by Hedges and Nowell (1995) for earlier decades, and the magnitude of gender differences remains stable over historical time. Additionally, we found a developmental effect, such that gender differences were found in much younger students but increased with additional years of schooling. International studies of reading achievement find that greater female performance is found universally across all nations (Guiso et al., 2008; Reilly, 2012), which would at least be consistent with biological factors. Yet there is also substantial variability across nations in the size of the gender gap, with sociocultural factors such as nation’s level of gender equality and gender norms also making a strong contribution (Reilly, 2015). Like many gender researchers, we advocate a broad biopsychosocial model of factors that contribute to gender differences (Halpern, 2000), rather than any single cause.

Despite the strengths that a large demographically representative sample like NAEP offers, the chief limitation is that it is limited to students from the United States, and to the way in which reading and writing skills are measured. Educational practices and frameworks clearly differ from country to country. International assessments of students’ reading achievement such as PIRLS and PISA (Lynn & Mik, 2009; Reilly, 2012) have found that the gender difference in reading is universal (i.e., all countries find girls significantly and meaningfully outperform boys). However, it is unclear whether a similarly sized gender difference in writing skills would exist internationally. Additionally due to limitations of the dataset (such as lack of subgroup sample sizes in the publicly available data), it was not possible to investigate Gender × Ethnicity interactions or socioeconomic status differences, a limitation shared by other analyses of NAEP data (e.g., Reilly et al., 2015). One factor that cannot be controlled in this dataset though is student dropout rates. As compulsory schooling extends in most U.S. states up to age 16, there should not be any meaningful attrition between Grades 4 and 8. While high school completion rates have been steadily increasing over the historical time period examined, the gender differences reported for Grade 12 do not include young adults that leave before this time (and presumably may have poorer reading and writing proficiency). As more girls complete high school than boys, this may underestimate the extent of gender differences in reading and writing in the general population. Additionally, it only presents a snapshot of students enrolled in U.S. schools—children that are unable to attend formal schooling due to other factors such as intellectual disability would likewise not be measured.

**Conclusion**

Gender differences in reading and writing achievement were found across all levels of the ability spectrum. Girls outperformed boys in mean reading and writing achievement, and contrary to our hypothesis these gender differences do not appear to be declining over the time period analyzed (1988–2015). Furthermore, there were pronounced differences in gender ratios for poor readers/writers, with boys greatly overrepresented. This pattern was reversed for those students attaining an advanced proficiency standard, with significantly more girls than boys. Our study also examined gender differences in younger students than those reported by Hedges and Nowell (1995), finding a developmental effect toward larger gaps as students progress through their schooling. These findings hold educational implications for students’ academic success during primary and high school, as well as academic readiness to embark on further college studies. A challenge for researchers is to identify the precise nature of gender differences in reading and writing so that educators can design targeted interventions to improve children’s reading and writing skills.

**References**


GENDER DIFFERENCES IN READING AND WRITING


Chapter 6 – Cross-Cultural Patterns of Reading, Mathematics and Science Literacy

This chapter reports on a meta-analysis of student testing data from the 2009 wave of the Programme for International Student Assessment (PISA), a large-scale educational assessment of student’s reading, mathematics and science literacy across all OECD members and a number of partner nations. This study reports data from 65 nations. Consistently across all nations, girls outperform boys in reading literacy, $d = -.44$. Boys outperform girls in mathematics in the USA, $d = +.22$ and across OECD nations, $d = +.13$. For science literacy, while the USA showed the largest gender difference across all OECD nations, $d = +.14$, gender differences across OECD nations were non-significant, and a small female advantage was found for non-OECD nations, $d = -.09$. Across all three domains, these differences were more pronounced at both tails of the distribution for low- and high-achievers. Considerable cross-cultural variability was also observed, and national gender differences were correlated with gender equity measures, economic prosperity, and Hofstede’s cultural dimension of power distance. Educational and societal implications of such gender gaps are addressed, as well as the mechanisms by which gender differences in cognitive abilities are culturally mediated.

It has been published as


Copyright statement:

It was published in accordance with the Creative Commons Attribution (CC_BY) license, and copyright was retained by the author.
Gender, Culture, and Sex-Typed Cognitive Abilities

David Reilly*
School of Applied Psychology, Griffith University, Southport, Queensland, Australia

Abstract

Although gender differences in cognitive abilities are frequently reported, the magnitude of these differences and whether they hold practical significance in the educational outcomes of boys and girls is highly debated. Furthermore, when gender gaps in reading, mathematics and science literacy are reported they are often attributed to innate, biological differences rather than social and cultural factors. Cross-cultural evidence may contribute to this debate, and this study reports national gender differences in reading, mathematics and science literacy from 65 nations participating in the 2009 round of the Programme for International Student Assessment (PISA). Consistently across all nations, girls outperform boys in reading literacy, \( d = -0.44 \). Boys outperform girls in mathematics in the USA, \( d = 0.22 \) and across OECD nations, \( d = 0.13 \). For science literacy, while the USA showed the largest gender difference across all OECD nations, \( d = 0.14 \), gender differences across OECD nations were non-significant, and a small female advantage was found for non-OECD nations, \( d = -0.09 \). Across all three domains, these differences were more pronounced at both tails of the distribution for low- and high-achievers. Considerable cross-cultural variability was also observed, and national gender differences were correlated with gender equity measures, economic prosperity, and Hofstede’s cultural dimension of power distance. Educational and societal implications of such gender gaps are addressed, as well as the mechanisms by which gender differences in cognitive abilities are culturally mediated.

Introduction

Rightly or wrongly, the topic of gender differences in cognitive abilities appears perennial, holding curiosity not only for social scientists but also for the general public and media [1–4]. Intelligence is multifaceted [5–10], and comprises a range of culturally-valued cognitive abilities. While there is almost unanimous consensus that men and women do not differ in general intelligence [11–14], there are several domains where either males or females as a group may show an advantage, such as visuospatial [15–16] and verbal abilities [17–18] respectively. However, gender differences in quantitative abilities [19], such as science and mathematics, remain contentious. Researchers are divided between arguing for small but still influential differences in quantitative reasoning [9–11], and claiming that any observed differences in maths are so small, in fact, that they can be categorised as ‘trivial’ [12–14].

A key limitation of research in this area is that it is largely US-centric, and does not speak to gender differences between males and females raised under different social and educational environments in other cultures. Additional lines of evidence are required, and one such source is international testing of students. Large national and international samples can provide a ‘yardstick’ estimate of gender differences within a given region, at a given point in time. By drawing from a broad population of students, national and international testing provide us with stronger evidence for gender similarities or differences than could be found from smaller, more selective samples. It is common practice for gender difference studies to use convenience samples drawn from psychology student subject pools [21], as well as from groups of high performing students such as gifted and talented programmes [22] – conclusions drawn from such samples may not be generalizable to wider populations. There is evidence to suggest that the performance of males is more widely distributed, with a greater numbers of high and low achievers [23]. This has been termed the greater male variability hypothesis [10,15–16], and presents a problem for researchers recruiting from only high achievers – even though mean differences between males and females may be equal, if
the distribution of male scores is wider than females, males will be overrepresented as high-achievers in a selective sample. This may lead to the erroneous conclusion that gender differences exist in the population of males and females.

A good example of this in practice comes in the form of the Scholastic Assessment Test (SAT) used for assessing suitability of students for college entry within the United States. Males consistently outperform females on the mathematical component [22,24–25]. Gender differences in SAT-M are extremely robust across decades, see Figure 1. On the basis of this evidence alone, one might erroneously conclude that the gender gap in mathematics is pervasive unless consideration is given to the demographics of the sample. Students considering college admission are motivated to undertake the SAT, and this is largely a self-selected sample that may differ on important characteristics such as socioeconomic status, and general ability level. Additionally many more girls sit the SAT than boys [24,26], reflecting the higher admission rate of women in college [27]. Thus the sample of males is more selective, while the sample of females is more general. One cannot rule out the possibility that the male sample includes a greater proportion of high achieving students and that the female sample may have included students of more mediocre mathematical ability, lowering mean performance.

This does not mean, necessarily, that one should discount any finding of gender differences in the SAT-M as being invalid. Data from the SAT may be extremely useful in estimating gender differences in the population of students considering further education. This is a very narrow, quite specific theoretical question. But such findings cannot be easily generalised to the general population, which is what researchers and laypersons alike would seek to test.

Another source of information on gender differences comes from experimental research carried out in the laboratory, under tightly controlled conditions. Equal numbers of males and females can be recruited using random selection. When large samples are randomly drawn from the general population, the scores of both high and low achievers are included in measurements of gender differences. Such studies are time-consuming and expensive to conduct, however. More commonly, gender difference studies use much smaller convenience samples, such as a subject pool of college students which also introduces the problem of selection bias [21]. College subject pools differ from the general population across many different characteristics [28], such as socioeconomic status, general intelligence, and prior educational experiences. Since the scores of males are more variable [12,18–19], a convenience sample that draws from only the upper-tail of ability will be skewed with a greater frequency of high performing males than females, thus exaggerating any gender difference that is found.

Additionally, many cognitive abilities show an interaction between gender and socioeconomic status [1,25–28]. Studies that selectively recruit from college subject pools in medium- to high- socioeconomic status regions would therefore be more likely to find gender differences than those recruiting from lower socioeconomic regions, as there will be greater differentiation between high and low ability levels. Likewise, samples drawing from a college pool may find greater gender differences than if they were recruited from a high school sample, or from the general population. Potentially, this could give a distorted picture of actual gender gaps when generalising from these selective samples to the wider population of males and females.

Large national samples allow researchers to investigate objectively the existence and magnitude of gender differences or similarities. We can be more confident that any observed differences are reflective of what we would find in the general population of boys and girls, and are not simply due to sampling bias. As additional waves of testing are conducted using similar measurement instruments, we can also begin to track any changes over time. It allows us to evaluate efforts aimed at reducing gender differences, and to see areas where further progress must be made. Such data may also be of benefit to policy makers and educational institutions in advocating for educational change, and in support of programs aimed at addressing inequalities.

Figure 1. Gender differences in SAT-M performance. On average, boys score higher than girls on the SAT-M exam (approximately one third of a standard deviation). The pattern of scores is consistent across years and does not appear to be diminishing, contrary to other lines of evidence that show gender differences in mathematics are small [51].
doi:10.1371/journal.pone.0039904.g001
Gender Differences in Mathematics and Science within the United States

For the United States, one such program is the National Assessment of Educational Progress (NAEP), a federal assessment of educational achievement. The NAEP is conducted for all states within the United States and since participation is both comprehensive and not self-selected, is ideally suited to answering the question of whether males and females differ in mathematical ability (a type of quantitative reasoning). Hyde [20] and colleagues examined gender differences between boys and girls in mathematics from grades 2 through 11, drawing on a sample of students from ten states which amounted to a sample of over seven million students. Hyde, et al. [20] reported an effect size for gender differences in each grade that approached zero, and categorised differences between males and females as “trivial” [29].

While this evidence seems quite compelling, one must be cautious about generalising the conclusion of ‘no difference’ in maths performance on the NAEP to maths performance in all areas of mathematics. As Hyde and Mertz [29] acknowledge, the test content of the NAEP does not include complex test items, making it impossible to investigate gender differences in this area. Complex and novel mathematical problem-solving is a prerequisite skill for success in many academic areas but most particularly in STEM-related fields. With increased affordability and access to calculators and computers, basic computation skills have become less important than the ability to understand complex problems and find strategies to solve them. A comprehensive meta-analysis conducted by Hyde, Fennema and Lamon [30] found small to medium sized differences in complex problem solving favoring males ($d = .29$). Assessment that includes these types of mathematical problems, therefore, should presumably show larger gender differences and might not necessarily support the gender similarities hypothesis. Evidence from the NAEP may exhibit a ceiling effect, as test content hasn’t adequately provided the opportunity for differentiation between high and low ability levels in complex reasoning. This would make the distribution of scores largely homogenous, preventing us from adequately testing the gender differences/gender similarities hypothesis.

International Sampling of Science and Mathematical Ability

Another source of evidence for evaluating claims of gender differences comes from international testing of students’ educational attainments as part of the OECD’s Programme for International Assessment (PISA). Beginning in 2000 and conducted every three years, participating nations assess the educational attainment of students using a standardized exam that allows their performance to be compared globally. PISA aims to assess the educational progress of students as they reach the end of compulsory education, at age 15, across three skill areas: these being reading literacy, mathematical literacy, and science literacy. Samples are stratified random probability samples, selected from a range of public and private institutions across geographical regions, and weighted so as to be nationally representative [31]. This overcomes the selection-bias of tests such as the SAT-M [24,26], as well as providing a more valid assessment of the general population of boys and girls at that age than could be found in college-bound students.

Additionally, the test content of PISA is somewhat different to that of other national testing assessments, such as the NAEP. PISA assesses both knowledge and problem-solving skills, reflecting the type of real-world content and skills required to be an informed and capable information consumer and citizen. It assesses a student’s reading, mathematical and scientific literacy, their ability to solve problems and to apply their knowledge and skills across each of these three domains. This is in contrast to tests that require primarily memory of learned material from the curriculum, allowing for greater differentiation between high and low ability levels. As such, it taps higher level cognitive skills than may be found in testing schemes like NAEP, which Hyde and colleagues have reported show small or trivial gender differences in science and mathematics [20]. The test content is sufficiently demanding that only 1.9% of US students are classified as attaining the highest proficiency level in mathematics, and only 1.3% of US students in science. While this makes it ideal for testing for gender differences or similarities within a given country such as the US, it also affords the opportunity to study them cross-culturally.

Cross-national Variation in Cognitive Abilities

Cross-national variation in the magnitude of gender differences can provide useful information about the environmental conditions that foster, or inhibit, gender differences in domains such as mathematics. While gender differences in mathematics are frequently found at a national level, they are not found universally across all nations [32]. Social roles for women vary greatly from culture to culture, with some cultures promoting higher standards of gender equality and access to education than others [33]. Even those nations that have progressive attitudes towards women may still have strongly-held cultural stereotypes that narrowly constrain them [34–38]. Cultural stereotypes that girls and women are less able than boys and men in mathematics and science still endure [39–40], and these stereotypes have damaging consequences for the self-efficacy of young girls [41].

Cross-cultural comparisons of the performance of males and females might help answer some theoretical questions about the origins of any observed gender differences. When we see consistent gender differences across many or all nations, and when they are large enough in magnitude to have a practical impact on the educational and occupational aspirations of boys and girls, then we might reasonably conclude some systematic process is responsible – be this biological or institutional. When we see changes in the magnitude and the direction of gender differences, as is the case for science performance reported below, then we might reasonably conclude that either cultural or environmental influences are strong moderators in the development of cognitive ability - gender differences are not an inevitable consequence of biology. Finally, if we were to see more similarities than differences in the performance of boys and girls, then this would also be useful information for shaping public policy and educational practices such as continuing support for coeducation [42].

A number of previous studies have examined the size of gender differences in cognitive abilities cross-culturally in an attempt to shed light on the underlying causes of such variation. Baker and Jones [43] reported strong correlations between measures of gender equity (such as percentage of females in higher education and the occupational status of women in society) and gender differences in mathematics. Gender differences in mathematics were smaller in more gender-equal nations than in less-equal nations. Though the precise mechanism by which this occurs is unclear, these findings have been replicated by a number of researchers [31–32,43]. This suggests that two factors influencing the cognitive abilities of women are the gender stereotypes that a culture holds, and the gender-roles for women in a society [29,32]. This has been referred to in the literature as the gender stratification hypothesis [33,43], which argues that gender differences are more pronounced when the
roles of men and women are tightly controlled into separate spheres and duties [35,37,44–45].

Mathematics is not the only cognitive domain where we see an influence of gender-equality and gender stereotypes on cognitive performance. The female advantage in reading and language, while universal, also differs in magnitude between nations. Guiso, et al. [32] examined data from the PISA 2003 round of testing, replicating the finding of Baker and Jones for mathematics as well as finding an association between gender equity and the gender gap in reading. Although this might be expected given that correlations between mathematics performance and reading overlap, the direction of the association differed. Instead of finding reduced gender differences in reading for countries fostering greater gender-equality, the gender gap between boys and girls actually increased. One possibility for this seemingly paradoxical finding is that whatever natural advantage girls may have for reading is suppressed in more restrictive countries, but that under favorable conditions is allowed to flourish to its full potential. However, further replication of these findings with subsequent waves of testing is required to determine whether this association is stable across time.

Programme for International Student Assessment (PISA) 2009

Cross-cultural evidence of gender differences or similarities provides a stronger foundation for understanding the role of social and biological factors in the development of sex differences, as noted above. The aim of this study was to explore sociocultural factors that promote, or inhibit, the development of gender gaps in highly sex-typed academic domains of reading, mathematics and science [46]. It presents findings from international assessment of student abilities as part of the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (OECD). The study uses data from the most recent round of testing to calculate national and international gender gaps in reading, mathematics, and science literacy.

In addition to presenting data on national gender differences, it uses meta-analytic techniques to calculate global gender differences to examine evidence for Hyde’s gender similarities hypothesis [47], which posits there are no meaningful gender differences in cognitive performance. The study also seeks to replicate the findings of past researchers for the gender stratification hypothesis [27,38,43–44], using several measures of gender equity and occupational segregation. A number of other sociocultural constructs are also examined to determine the extent to which gender differences are culturally mediated by factors other than biology.

One hypothesised influence is the economic prosperity of a nation [39–41], which reflects two mechanisms. Firstly, greater economic prosperity allows for a greater proportion of national resources to be spent on education, resulting in a higher quality of education and emphasis on skills such as mathematics and science. Secondly, skills in these technical areas are in greater demand, and represent a pathway to a higher standard of living. This may result in greater competition for these occupations, and such competition may not always be helpful to the career aspirations of women wishing to enter male-dominated fields. While increases in gender equity are strongly associated with economic prosperity (and hence should be associated with smaller gender gaps), these may be partially offset by increased occupational stratification and stronger cultural stereotypes associating maths and science with gender roles [27,32–33,44–45]. Thus increased gender differences are not purely the result of increased spending on education and also reflect social processes.

A second mechanism by which gender differences may be culturally mediated is through the attitudes, values, and beliefs of a nation. While beliefs about the role of women in society vary considerably from nation to nation, there are few instruments available that have wide global coverage of gender stereotypes and attitudes towards women [38,48–49]. One of most widely used cultural instruments is Hofstede’s [50] five cultural dimensions. One of these is theoretically relevant to cultural mediation of gender differences in cognitive ability, the dimension of power distance.

The dimension power distance describes the ways in which societies address the issue of human inequality, and the ways in which social groups are segregated [50]. In a lower power distance culture, there are reduced distinctions between social classes, between employees and employers, between students and teachers, and between genders. Higher power distance cultures have greater social division, and a compensatory strategy for those who are lower in power is to acquire culturally valued skills through education. Girls may have increased motivation to learn maths and science and pursue higher status occupations as a way of overcoming social inequity.

Hypotheses

Based on prior research and theoretical perspectives, it was hypothesised that:

1) Gender differences in the domains of mathematics, and science would be found for the United States, and these would be larger than those reported by Hyde [51]. These would reflect gender stereotypes associating these domains with masculinity and males [39]. However gender differences cross-culturally would be much smaller, in partial support of a global gender similarities hypothesis.

2) Gender differences in reading performance in favor of girls would be found in reading for the United States and cross-culturally, reflecting an inherent biological disposition that is only weakly influenced by cultural environment.

3) Measures of national gender equity would be associated with smaller gender gaps in mathematics and science, in support of the gender stratification hypothesis. Furthermore, increased gender equity would be weakly associated with wider reading gaps in favor of girls.

4) Economic prosperity would be associated with wider gender gaps in mathematics and science than in less prosperous nations, reflecting increased spending on education, increased demand for these skills, and heightened competition by males. Such competition may not be helpful to the career aspirations of women, but will not influence reading performance which is less malleable to social and cultural influences.

5) Countries that score highly on Hofstede’s power distance dimension have greater segregation and foster inequalities, particularly for women. A compensatory strategy for women is to acquire culturally-valued skills such as science and mathematics. High power distance nations would be associated with smaller gender gaps or a slight female advantage in these domains. Boys may have increased motivation to develop reading and writing proficiency in high power distance cultures, resulting in smaller gender gaps for reading literacy.
Methods

Participants

Performance data for students accessed under PISA is offered as a publicly accessible archive for researchers. Additionally, aggregate national performance profiles are published as separate male and female subgroups [31], which were used for analysis. PISA 2009 included 34 OECD countries, as well as 31 additional partner nations. This amounts to a total participant size of 430,405 students (50.6% female) drawn from across 65 nations. This represents the most recent round of testing, as well as providing performance data for a broader range of nations than earlier PISA assessments.

Analysis

National performance profiles in reading, mathematics and science literacy were obtained from OECD [31], which reports the assessment of boys and girls separately. Because of the large sample sizes involved in national testing, even slight or trivial differences between boys and girls may be deemed statistically significant, even though it may have no practical significance. For this reason, an effect size is presented in the form of Cohen’s $d$, the mean standardized difference. This allows the reader to draw his or her own conclusions as to the practical significance of reported gender differences.

The computation is calculated as the mean difference between male and female scores, divided by the pooled within-gender standard deviation. By convention, female scores are subtracted from male scores, so that a positive $d$ indicates higher scores for males while a negative $d$ reflects higher scores for females. This convention is observed for readability reasons only, and the interested reader may choose to rephrase the equations so that male scores are subtracted from female scores simply by inverting the sign of any effect size given.

Conventional criteria for labelling effect sizes as "small", "medium", or "large" have many limitations and should be used with great caution [52–53]. Cohen [53] offered a rule of thumb that an effect size of $d \leq 0.20$ could be considered a "small" effect for the purpose of estimating statistical power, and that many legitimate psychological phenomena studied are in fact small effects. The label of small is perhaps an unfortunate one as some researchers have mistakenly taken small to be of no practical significance, a practice Rosenthal and Rubin [54] caution against. However Hyde, et al. [20] have argued that effect sizes as small as $d = 0.04$ should be regarded as trivial, a cut-off which seems sound practice. Hyde [47] has also suggested that $d \leq 10$ should be actually be regarded “as close to zero” (p.581), a cut-off which is overly conservative and dismisses what are legitimate, albeit very small, between-group differences. Accordingly, Cohen’s conventions for labelling are followed for reporting. Additionally, gender differences are presented using Rosenthal and Rubin’s [54–55] Binomial Effect Size Display (BESD) which presents results in a metric that represents effect size in a format suitable for interpretation by non-statisticians [56].

In order to test the gender similarities hypothesis, national gender gaps in reading, mathematics, and science were combined using meta-analysis. Comprehensive Meta Analysis (CMA) V2 software was used for the calculation of statistics [57]. A random-effects model was chosen [58] due to the high degree of cross-cultural variability, which would make a fixed-effects model unsuitable [56,59]. Such a method is more conservative in estimating error terms and produces wider confidence intervals, giving us greater assurance that the true effect size falls within this range.

Favreau [60] argues against the use of null hypothesis testing for evaluating claims of gender difference because it may be overly sensitive, and does not present a clear picture of how differences are distributed across groups. Accordingly, data is presented showing high and low-achievers, as well as effect sizes. Even when a mean gender difference may be regarded as ‘small’ by Cohen’s [53] conventions, or ‘trivial’ by Hyde [47], a more pronounced difference may be found at the tails of a distribution in high and low-achieving students, resulting in quite disparate educational outcomes.

Moderation effects of sociocultural factors were examined to test the gender stratification hypothesis for national gender gaps using correlational analysis. Although past researchers [32,61] have examined the gender stratification hypothesis for mathematics and reading, exploration of the relationship with science has gone largely untested. Multiple measures of gender equity were used, as each instrument operationalises the construct of gender equity differently, and prior research has shown that they vary in their predictive validity for educational and social outcomes. Other moderators tested include economic prosperity, as measured by GDP, and Hofstede’s power distance dimension.

Gender gap index. For comparability with Guiso, et al.’s findings, the Gender Gap Index (GGI) produced by the World Economic Forum was selected as one measure of gender equity [62]. Data for the calendar year of PISA testing was used. This measure assesses four areas: economic participation, educational attainment, political empowerment, and health and survival. While the first three are theoretical relevant to the gender stratification hypothesis, health and survival (which measures differences in male and female life expectancy, as well as sex ratio) may reflect other - largely biological – factors, thus lowering predictive validity of this measure. An additional criticism of this measure is that the economic participation component emphasises male to female participation across various sectors, but gives less emphasis to income disparities.

Relative status of women. As an alternative conceptualisation of gender equity, the Relative Status of Women (RSW) measures gender differences across educational attainment, life expectancy, and women’s share of income [63]. This reflects a stronger economic and educational component in estimation of gender stratification, with wage inequality playing a greater weighting.

Women in research. Else-Quest, et al. [61] argued that domain-specific indicators of gender equity may play an important role in the development of gender differences, with those related to gender stratification in educational outcomes showing strong predictive validity. One such marker is the relative share of research positions held by women. Data for this measure was obtained from the UNESCO Institute for Statistics, and supplemented by data from the National Science Foundation and Statistics Canada. Data was selected for the calendar year 2009 when possible, or earlier if not available. Women’s relative share of research positions was available for forty one nations.

Gross domestic product (GDP). Economic data was obtained from the World Economic Outlook database produces by the International Monetary Fund. Archived information for the calendar year 2009 was obtained for sixty-one nations.

Hofstede’s power distance index. National power distance scores are published in Hofstede’s text “Culture’s consequences” [50], which ranks nations across this dimension. Data was unavailable however for many of the non-OECD partner nations, and several European countries, and was supplemented by national profiles published online (http://geert-hofstede.com). This provided coverage of fifty two nations.
Statistical Power

While the sample size represented by the PISA 2009 was extremely large, when examining gender differences at the country level (n = 65) for correlation analysis the sample size is relatively small. Additionally, data for gender equity measures and for Hofstede’s cultural dimension of power distance was unavailable for many non-OECD nations reducing sample size even further. With a reduced sample size, correlations may lack sufficient power to detect relationships that are relatively weak in nature [53]. Given that hypotheses were directional (e.g. greater gender equality would be associated with a reduced gender gap in mathematics), a decision to make correlation tests one-tailed would often have allowed such a correlation to be deemed statistically significant (as probability values are halved). For this reason exact probability values are given, along with the size of the correlation coefficient, so that the reader can decide whether to make the appropriate adjustment. All tests report two-tailed correlations unless otherwise specified. Data for one nation, Colombia, represented both a univariate and multivariate outlier, and was excluded from all correlational analysis.

Results

Although assessing qualitatively different abilities, there was a strong overlap between national gender differences in reading, mathematics and science. The quantitative abilities of mathematics and science showed the greatest overlap. Table 1 presents intercorrelations between national gender differences in these domains, while Table 2 gives correlations between national predictor variables. Tables 3 and 4 present national sample size and calculated effect sizes across the three domains for OECD and partner nations respectively.

Reading Literacy

Table 5 presents summary statistics for reading achievement. Within the United States, girls outperformed boys in overall reading. Cohen’s $d = - .26$ which is just over a quarter of a standard deviation. By comparison, the OECD gender difference in reading was larger, $d = - .42$. Examining performance data for the US sample further, boys were overrepresented at the lowest level of reading proficiency, with approximately 4.5 boys to every girl. When we consider the vocational and economic outcomes associated with poor literacy, such a disparity is alarming. Such findings are consistent with previous findings on reading literacy assessed by PISA [32] and gender differences in the prevalence of reading difficulties [64–65]. When we look at students attaining the highest level of reading proficiency (Level 6), the trend is reversed with over twice the number of girls than boys achieving the highest standard. Thus boys are overrepresented at the lower end of the spectrum, while girls are overrepresented at the highest end.

Overall, across all sixty-five nations the gender difference in reading literacy favored girls, $d = -.44$ [95%CI = -.41, -.46], $Z_{max} = -31.04$, $p < .001$, with a similar gender difference also being found for OECD nations only as a group. Additionally, statistically significant gender differences in reading favoring girls were found in every nation surveyed, and have since the first assessment in 2000 [66]. These effect sizes ranged from $- .11$ to $- .68$, from a small- to a medium- sized difference in reading literacy.

To investigate the gender stratification hypothesis, I examined correlations between gender equity and the gender gap in reading. Partial support was found for the gender stratification hypothesis. National scores on the Relative Status of Women (RSW) measure were negatively correlated with reading, $r = -.33$, $p = .018$, such that increased gender equality was associated with larger reading gaps favoring females. Additionally, the educational measure of women in research (WIR) was associated with larger reading gaps, $r = -.38$, $p = .016$. Surprisingly though, there was no association between the gender gap index (GGI) and reading ability, $r = .01$. Examination of the scatterplot showed no discernable pattern, and the result was not driven by outliers.

Stronger support for the gender stratification hypothesis was found when examining gender differences in the percentage of students attaining the highest level of reading. Improvements in national gender equity was associated with a wider gender gap in high achieving girls, RSW, $r = -.32$, $p = .021$; GGI, $r = -.41$, $p = .002$, which is consistent with the findings of Guiso, et al. [32]. Somewhat surprisingly, however, the educational measure of gender equity showed a strong positive association, with increases in the percentage of women in research associated with smaller gender gaps, $r = .57$, $p < .001$. While the role of women in higher education may make a contribution to the mean performance of girls and boys in basic reading literacy, it may be the case that for high-achieving reading comprehension skills, boys and girls benefit equally from female role-models in higher learning.

No association between GDP and gender differences in reading was found, $r = .04$, consistent with predictions. However, a strong association with economic prosperity was found for reading high achievers, $r = -.43$, $p < .001$ with a greater ratio of female to male high achievers as GDP increased. This suggests an interaction between gender and GDP, with girls benefiting more from economic prosperity than boys. Furthermore, while no association was found between power distance and mean reading literacy scores of boys and girls, a strong positive association with the gender gap in high achievers was found as hypothesized, $r = .40$, $p = .003$ with gender ratios approaching more equal representation as power distance increased. Cultural mediation through economic prosperity and power distance was not found for mean male and female performance, only for gender ratios in high achievement.

Mathematics Literacy

Table 6 presents summary statistics for mathematics literacy. Within the United States, boys scored higher on mathematical literacy than girls, $d = .22$ which is a small but non-trivial effect size. Additionally, the size of the gender differences was almost twice that of the OECD average. This is in contrast to previous studies examining national mathematics performance by Hyde, et al. [20] which had found a gender gap that approached zero. At the lower end of ability level for the US sample, the difference in prevalence between girls and boys was extremely slight; however at the highest ability levels there were just over twice as many boys than girls reaching this proficiency level.

Table 1. Correlations between National Gender Differences for PISA Reading, Mathematics, and Science Performance (All Nations).}

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Mathematics</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>1.00</td>
<td>.75***</td>
<td>.78***</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1.00</td>
<td>.81***</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

*p < .05,
**p < .01,
***p < .001.

doi:10.1371/journal.pone.0039904.t001
As the distribution of gender differences differed somewhat between OECD and partner nations, they are reported separately. Overall, across all 34 OECD nations, there was a significant gender difference favoring males on mathematical literacy, Cohen’s $d = .13$ 95%CI $[.11,.15]$, $z_{ma} = 11.22$, $p < .001$. While this is a small effect size, it does exceed the criteria set forth by Hyde and Linn [51] for trivial gender differences. Gender differences across PISA partner nations also favored males, Cohen’s $d = .07$ 95%CI $[.02,.11]$, $z_{ma} = 3.10$, $p = .001$ although this difference was somewhat smaller.

While statistically significant differences were found in most countries, they showed considerable variability ranging from $d = -.12$ to $d = .43$ (see Figure 2). For many nations the gender gap is negligible, while others show small to medium sized differences. Additionally the direction of the gender gap was sometimes reversed, with girls outperforming boys in many nations. Under different social and educational environments, a gender advantage supporting either males or females emerges. This would be inconsistent with Hyde’s [47] gender similarities hypothesis; rather, gender differences or similarities in mathematics are strongly mediated by cultural factors.

To explore the gender stratification hypothesis, correlations between gender equity measures and the gender gap in maths were examined. As hypothesized there was a strong negative relationship between the educational measure of women in research and the gender gap in mathematics, $r = -.38$, $p = .014$. Greater representation of women in research was associated with smaller gender gaps or a female advantage, consistent with the findings of Else-Quest, et al. [61]. However, only a weak association was found between gender equity measure of RSW, $r = -.14$, and no association was found between GGI and maths, in contrast to the findings of Guiso, et al. [32].

Since the PISA 2009 dataset includes a much broader range of partner nations than was examined by Guiso, et al. [32], the strength of the gender equity association may have been obscured by additional noise reflecting developed/developing nationhood. When restricting analysis to OECD nations only, the hypothesized gender equity association was found for the relative status of women (RSW) measure, $r = -.42$, $p = .020$, as well as a weak association with GGI, $r = -.21$ that fell short of statistical significance. While gender equity plays an important role in the development of gender differences in mathematical literacy for developed nations, it may be the case that there are more proximate needs for girls in developing nations (such as access to schooling, parental support, freedom from work and home duties) that these gender equity measures do not assess.

A similar pattern of associations was found for gender differences in high achieving mathematics students across all nations. There was a strong association between women in research educational measure, $r = -.63$, $p < .001$, with increased representation of women in research positions associated with a smaller gender difference in high achievers approaching zero (see Figure 3). However no association was found between the gender gap in high achievers and other gender equity measures, nor was this found when restricting to OECD nations only.

Support was also found for the economic prosperity hypothesis. Mean gender differences in mathematics literacy were larger in more economically prosperous nations, $r = .31$, $p = .015$. This relationship was stronger for high achievement, $r = .53$, $p < .001$ with a greater number of males attaining this level of proficiency.

Examining the relationship between Hofsetede’s power distance cultural dimension and mathematics literacy, support was also found for cultural mediation. There was a strong negative relationship between power distance and mean gender differences in mathematics, $r = -.28$, $p = .044$, as well as for gender ratios in high achievement, $r = -.33$, $p = .019$. Gender differences were smaller in nations with greater tolerance for inequality, suggesting a compensatory strategy to acquire culturally and economically valued skills in mathematics.

### Table 2. Correlations between Measures of Gender Equity, Economic Prosperity, and Hofstede’s Power Distance Index.

<table>
<thead>
<tr>
<th></th>
<th>Gender Gap Index (GGI)</th>
<th>Relative Status of Women (RSW)</th>
<th>Relative Share of Women in Research (WIR)</th>
<th>Gross Domestic Product (GDP) per capita, 2009</th>
<th>Hofstede’s Power Distance Index (PDI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGI</td>
<td>1.00</td>
<td>.43**</td>
<td>−.09</td>
<td>.43**</td>
<td>−.59***</td>
</tr>
<tr>
<td>RSW</td>
<td>1.00</td>
<td>−.11</td>
<td>.05</td>
<td>−.38*</td>
<td>1.00</td>
</tr>
<tr>
<td>WIR</td>
<td>1.00</td>
<td>−.60***</td>
<td>−.20</td>
<td>−.58***</td>
<td>1.00</td>
</tr>
<tr>
<td>GDP</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

*p < .05,
**p < .01,
***p < .001.

[doi:10.1371/journal.pone.0039904.t002]
Gender similarities, rather than differences, were the norm which is consistent with the findings of Hyde and Linn [51]. Somewhat surprisingly, there were also five nations where girls outperformed boys to a statistically significant degree (the largest being Finland, $d = -0.17$). One of the advantages of cross-cultural comparisons in national testing is that it highlights just how powerfully cultural and environmental influences can be in either promoting - or inhibiting - the cognitive development and learning of a child.

A markedly different picture of gender differences in science can be found across the 31 non-OECD nations. In general, females scored higher in science literacy than males across most nations. Overall, across non-OECD nations surveyed there was a statistically significant difference in science literacy favoring girls, $d = -0.09$ 95%CI $[-0.14, -0.04]$, $z_{na} = -3.44$, $p = .001$. For some nations, the gender difference was trivial or favored boys, but these were the exception; this is in contrast to the gender similarities in science noted above for OECD nations.

When both OECD and non-OECD nations were combined, there was a statistically significant difference in favor of girls, $d = -0.04$ 95%CI $[-0.07, -0.01]$, $z_{na} = -2.84$, $p = .005$. This effect size would fall into the trivial size by Hyde’s [67] conventions, but a focus on the combined sample overlooks the pattern of gender differences at a national level where girls show small but meaningful gains over boys in science literacy across large parts of the world. Given that women are underrepresented in science, particularly in the United States [68] such findings call into question the validity of cultural stereotypes that associate science

### Table 3. National Gender Differences in Reading, Mathematics, and Science Literacy forCountries within the OECD.

<table>
<thead>
<tr>
<th>Country</th>
<th>Sample size</th>
<th>Effect sizes (Cohen’s $d$)</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>7020</td>
<td>7231</td>
<td>-0.37</td>
<td>0.11</td>
<td>-0.01</td>
</tr>
<tr>
<td>Austria</td>
<td>3252</td>
<td>3338</td>
<td>-0.41</td>
<td>0.20</td>
<td>0.08</td>
</tr>
<tr>
<td>Belgium</td>
<td>4352</td>
<td>4156</td>
<td>-0.27</td>
<td>0.21</td>
<td>0.06</td>
</tr>
<tr>
<td>Canada</td>
<td>11431</td>
<td>11776</td>
<td>-0.38</td>
<td>0.14</td>
<td>0.05</td>
</tr>
<tr>
<td>Chile</td>
<td>2870</td>
<td>2799</td>
<td>-0.27</td>
<td>0.26</td>
<td>0.11</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>3115</td>
<td>2949</td>
<td>-0.53</td>
<td>0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>Denmark</td>
<td>2886</td>
<td>3038</td>
<td>-0.34</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>Estonia</td>
<td>2430</td>
<td>2297</td>
<td>-0.53</td>
<td>0.11</td>
<td>-0.01</td>
</tr>
<tr>
<td>Finland</td>
<td>2856</td>
<td>2954</td>
<td>-0.64</td>
<td>0.03</td>
<td>-0.17</td>
</tr>
<tr>
<td>France</td>
<td>2087</td>
<td>2211</td>
<td>-0.38</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>Germany</td>
<td>2545</td>
<td>2434</td>
<td>-0.42</td>
<td>0.16</td>
<td>0.05</td>
</tr>
<tr>
<td>Greece</td>
<td>2412</td>
<td>2557</td>
<td>-0.50</td>
<td>0.15</td>
<td>-0.11</td>
</tr>
<tr>
<td>Hungary</td>
<td>2294</td>
<td>2311</td>
<td>-0.42</td>
<td>0.13</td>
<td>0.00</td>
</tr>
<tr>
<td>Iceland</td>
<td>1792</td>
<td>1854</td>
<td>-0.46</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>Ireland</td>
<td>1973</td>
<td>1964</td>
<td>-0.41</td>
<td>0.09</td>
<td>-0.03</td>
</tr>
<tr>
<td>Israel</td>
<td>2648</td>
<td>3113</td>
<td>-0.38</td>
<td>0.08</td>
<td>-0.03</td>
</tr>
<tr>
<td>Italy</td>
<td>15696</td>
<td>15209</td>
<td>-0.48</td>
<td>0.16</td>
<td>-0.02</td>
</tr>
<tr>
<td>Japan</td>
<td>3126</td>
<td>2962</td>
<td>-0.39</td>
<td>0.10</td>
<td>-0.12</td>
</tr>
<tr>
<td>Korea</td>
<td>2590</td>
<td>2399</td>
<td>-0.45</td>
<td>0.04</td>
<td>-0.03</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>2319</td>
<td>2303</td>
<td>-0.38</td>
<td>0.20</td>
<td>0.07</td>
</tr>
<tr>
<td>Mexico</td>
<td>18209</td>
<td>20041</td>
<td>-0.29</td>
<td>0.17</td>
<td>0.08</td>
</tr>
<tr>
<td>Netherlands</td>
<td>2348</td>
<td>2412</td>
<td>-0.27</td>
<td>0.19</td>
<td>0.04</td>
</tr>
<tr>
<td>New Zealand</td>
<td>2396</td>
<td>2247</td>
<td>-0.44</td>
<td>0.08</td>
<td>-0.06</td>
</tr>
<tr>
<td>Norway</td>
<td>2375</td>
<td>2285</td>
<td>-0.52</td>
<td>0.06</td>
<td>-0.04</td>
</tr>
<tr>
<td>Poland</td>
<td>2443</td>
<td>2474</td>
<td>-0.56</td>
<td>0.04</td>
<td>-0.07</td>
</tr>
<tr>
<td>Portugal</td>
<td>3020</td>
<td>3278</td>
<td>-0.44</td>
<td>0.13</td>
<td>-0.04</td>
</tr>
<tr>
<td>Slovak Republic</td>
<td>2238</td>
<td>2317</td>
<td>-0.57</td>
<td>0.03</td>
<td>-0.01</td>
</tr>
<tr>
<td>Slovenia</td>
<td>3333</td>
<td>2822</td>
<td>-0.60</td>
<td>0.01</td>
<td>-0.15</td>
</tr>
<tr>
<td>Spain</td>
<td>13141</td>
<td>12746</td>
<td>-0.33</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>Sweden</td>
<td>2311</td>
<td>2256</td>
<td>-0.46</td>
<td>-0.02</td>
<td>-0.04</td>
</tr>
<tr>
<td>Switzerland</td>
<td>6020</td>
<td>5790</td>
<td>-0.42</td>
<td>0.20</td>
<td>0.08</td>
</tr>
<tr>
<td>Turkey</td>
<td>2551</td>
<td>2445</td>
<td>-0.52</td>
<td>0.12</td>
<td>-0.15</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>6062</td>
<td>6117</td>
<td>-0.26</td>
<td>0.23</td>
<td>0.10</td>
</tr>
<tr>
<td>United States</td>
<td>2687</td>
<td>2546</td>
<td>-0.26</td>
<td>0.22</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Note: Significant gender differences are highlighted in bold.

### Table 4. National Gender Differences in Reading, Mathematics, and Science Literacy for PISA Partner Countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Sample size</th>
<th>Effect sizes (Cohen’s $d$)</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albania</td>
<td>2321</td>
<td>2275</td>
<td>-0.62</td>
<td>-0.12</td>
<td>-0.33</td>
</tr>
<tr>
<td>Argentina</td>
<td>2183</td>
<td>2591</td>
<td>-0.34</td>
<td>0.11</td>
<td>-0.08</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>2443</td>
<td>2248</td>
<td>-0.31</td>
<td>0.13</td>
<td>-0.10</td>
</tr>
<tr>
<td>Brazil</td>
<td>9101</td>
<td>11026</td>
<td>-0.30</td>
<td>0.19</td>
<td>0.04</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>2231</td>
<td>2276</td>
<td>-0.54</td>
<td>-0.04</td>
<td>-0.19</td>
</tr>
<tr>
<td>Colombia</td>
<td>3711</td>
<td>4210</td>
<td>-0.11</td>
<td>0.43</td>
<td>0.26</td>
</tr>
<tr>
<td>Croatia</td>
<td>2653</td>
<td>2341</td>
<td>-0.58</td>
<td>0.12</td>
<td>-0.10</td>
</tr>
<tr>
<td>Dubai (UAE)</td>
<td>5554</td>
<td>5313</td>
<td>-0.47</td>
<td>0.02</td>
<td>-0.26</td>
</tr>
<tr>
<td>Hong Kong-China</td>
<td>2257</td>
<td>2280</td>
<td>-0.39</td>
<td>0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>Indonesia</td>
<td>2534</td>
<td>2602</td>
<td>-0.55</td>
<td>-0.02</td>
<td>-0.13</td>
</tr>
<tr>
<td>Jordan</td>
<td>3120</td>
<td>3366</td>
<td>-0.63</td>
<td>-0.01</td>
<td>-0.39</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>2723</td>
<td>2689</td>
<td>-0.47</td>
<td>-0.01</td>
<td>-0.10</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>2381</td>
<td>2605</td>
<td>-0.54</td>
<td>-0.07</td>
<td>-0.24</td>
</tr>
<tr>
<td>Latvia</td>
<td>2175</td>
<td>2327</td>
<td>-0.59</td>
<td>0.02</td>
<td>-0.09</td>
</tr>
<tr>
<td>Liechtenstein*</td>
<td>181</td>
<td>148</td>
<td>-0.39</td>
<td>0.28</td>
<td>0.18</td>
</tr>
<tr>
<td>Lithuania</td>
<td>2287</td>
<td>2241</td>
<td>-0.68</td>
<td>-0.07</td>
<td>-0.20</td>
</tr>
<tr>
<td>Macao-China</td>
<td>3011</td>
<td>2941</td>
<td>-0.45</td>
<td>0.13</td>
<td>-0.03</td>
</tr>
<tr>
<td>Montenegro</td>
<td>2443</td>
<td>2382</td>
<td>-0.57</td>
<td>0.14</td>
<td>-0.14</td>
</tr>
<tr>
<td>Panama</td>
<td>1936</td>
<td>2033</td>
<td>-0.33</td>
<td>0.06</td>
<td>-0.02</td>
</tr>
<tr>
<td>Peru</td>
<td>3000</td>
<td>2985</td>
<td>-0.23</td>
<td>0.20</td>
<td>0.05</td>
</tr>
<tr>
<td>Qatar</td>
<td>4410</td>
<td>4568</td>
<td>-0.44</td>
<td>-0.05</td>
<td>-0.25</td>
</tr>
<tr>
<td>Romania</td>
<td>2378</td>
<td>2398</td>
<td>-0.47</td>
<td>0.04</td>
<td>-0.13</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>2633</td>
<td>2685</td>
<td>-0.50</td>
<td>0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>Serbia</td>
<td>2680</td>
<td>2843</td>
<td>-0.47</td>
<td>0.13</td>
<td>-0.01</td>
</tr>
<tr>
<td>Singapore</td>
<td>2358</td>
<td>2587</td>
<td>-0.50</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>Chinese Taipei</td>
<td>2911</td>
<td>2920</td>
<td>-0.43</td>
<td>0.05</td>
<td>-0.01</td>
</tr>
<tr>
<td>Thailand</td>
<td>2681</td>
<td>3544</td>
<td>-0.52</td>
<td>0.05</td>
<td>-0.16</td>
</tr>
<tr>
<td>Trinidad and Tobago</td>
<td>2283</td>
<td>2495</td>
<td>-0.51</td>
<td>-0.08</td>
<td>-0.17</td>
</tr>
<tr>
<td>Tunisia</td>
<td>2359</td>
<td>2596</td>
<td>-0.37</td>
<td>0.16</td>
<td>0.01</td>
</tr>
<tr>
<td>Uruguay</td>
<td>2810</td>
<td>3147</td>
<td>-0.42</td>
<td>0.13</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

Note: Significant gender differences are highlighted in bold.

*Although effect sizes are large, caution must be taken interpreting due to small sample size.

Note: Significant gender differences are highlighted in bold.

doi:10.1371/journal.pone.039904.0004
with masculinity [69], and highlight the need for further efforts at challenging these damaging cultural stereotypes.

Examining mean gender differences in science literacy, partial support for the gender stratification hypothesis was found. There was a strong correlation between national GGI scores and science, \( r = .29, p = .035 \), with greater gender equity associated with smaller gender gaps approaching zero. However, only a weak non-significant association was found for the RSW, \( r = .14 \).

Additionally, there was a strong negative correlation between the percentage of women in research and gender gaps in science, \( r = -.39, p = .011 \), with increased representation of women being associated with a stronger female advantage over males in science. Thus increased gender equity was associated with more equal science performance, but this was offset by higher female performance as the share of women in research positions increased.

Only weak support for the gender stratification hypothesis was found for gender differences in high achievement in science. Increased gender equity as measured by the percentage of researchers who are women was associated with smaller gender gaps in the number of high achievers, \( r = -.57, p < .001 \) (see Figure 5). While positive female role models are certainly important for challenging gender stereotypes about women in science generally, they may be even more so for encouraging young women to excel in science and pursue it as a career path. In contrast to this finding, there was no association between the relative status of women measure, \( r = .12 \) and a slight positive correlation with gender equity as measured by the GGI, \( r = .29, p = .029 \), with increased gender equity associated with more male high achievers than female which is contrary to predictions. This anomalous association may be at least partly explained by the underlying construct measured by the GGI. It incorporates a strong economic component in its formula, with a correlation of \( r = .43 \) between national GGI scores and economic productivity as measured by GDP. When controlling for economic productivity, the association between GGI and science high achievers becomes non-significant, \( r = .12, p = .373 \).

Strong support was also found for culturally mediated of gender differences in science. Positive relationships were observed between GDP and gender differences in mean science scores, \( r = .42, p = .001 \), as well as for gender ratios in high achievement, \( r = .27, p = .036 \), as hypothesised. In contrast, a negative relationship was found between the power-distance dimension and mean gender differences in science, \( r = -.39, p = .005 \) with gender differences favoring girls in high power-distance nations. This effect was even stronger for gender ratios in high achievement, \( r = -.45, p = .001 \).

**Discussion**

Does the size of gender differences in reading, mathematics, and science from PISA assessment merit further research into the social and cultural factors that promote, or inhibit, differential educational outcomes for boys and girls? Evidence presented for the United States shows that there are meaningful gender gaps across all three domains. Furthermore, they are larger than those found in most OECD nations placing the US among the highest gender gaps in mathematics and science in the developed world, but somewhat smaller than other nations in reading literacy. However, quite different patterns are found when examining gender gaps globally. US performance is reviewed first, followed by a discussion of cross-cultural evidence.

**Reading Literacy**

While a small-to-medium sized gender difference in reading was found for US students \( d = -.26 \), this was comparatively smaller than that found in other OECD nations. However, gender differences were strikingly different at both tails of the distribution, with boys overrepresented in the lowest level of reading proficiency and girls overrepresented in the highest. PISA sampling allows for exclusion of students with limited language proficiency, so it is likely that this result reflects poorer reading ability generally rather than male overrepresentation in reading difficulties students. This pattern is consistent with existing research on gender ratios for reading difficulties [64–65].

Cross-culturally, a medium sized gender difference \( (d = -.44) \) was found for reading literacy, which would be inconsistent with Hyde’s gender similarities hypothesis [47]. Expressed in the BESD

**Table 5. Reading Ability for Girls and Boys for the USA and OECD nations.**

<table>
<thead>
<tr>
<th></th>
<th>Girls</th>
<th>Boys</th>
<th>Standard Deviation</th>
<th>Effect Size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>513</td>
<td>488</td>
<td>(97)</td>
<td>-.26</td>
</tr>
<tr>
<td>OECD Average</td>
<td>513</td>
<td>474</td>
<td>(93)</td>
<td>-.42</td>
</tr>
<tr>
<td>% students at lowest ability level, USA</td>
<td>0.2%</td>
<td>0.9%</td>
<td>4.5 boys : 1 girl</td>
<td></td>
</tr>
<tr>
<td>% at highest ability level, USA</td>
<td>2.1%</td>
<td>0.9%</td>
<td>2.4 girls : 1 boy</td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0039904.t005

**Table 6. Mean Mathematical Ability for Girls and Boys for the USA and OECD nations.**

<table>
<thead>
<tr>
<th></th>
<th>Girls</th>
<th>Boys</th>
<th>Standard Deviation</th>
<th>Effect Size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>477</td>
<td>497</td>
<td>(91)</td>
<td>.22</td>
</tr>
<tr>
<td>OECD Average</td>
<td>490</td>
<td>501</td>
<td>(92)</td>
<td>.12</td>
</tr>
<tr>
<td>% students at lowest ability level, USA</td>
<td>9.5%</td>
<td>6.8%</td>
<td>1.4 girls : 1 boy</td>
<td></td>
</tr>
<tr>
<td>% at highest ability level, USA</td>
<td>1.2%</td>
<td>2.5%</td>
<td>2.12 boys : 1 girl</td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0039904.t006
format, the likelihood of being average or higher in reading ability increases from 39% for boys to 61% for girls. Reading performance was higher for girls than boys across every nation, but also showed considerable between-nation variation. Though the direction of gender differences would be consistent with a biological explanation, it appears at least partially malleable by social and cultural factors. While there was no support for cultural mediation through economic prosperity and power distance in mean gender differences, contrary to predictions associations were found for high achievers in reading literacy.

It has been a common research finding that boys are generally poorer readers and writers than girls [70], and considerable effort has been made to address the gender gap over recent decades with focus on early identification and intervention for reading difficulties. Basic literacy is an essential life skill for all children, and for full participation as a citizen. While much attention is given to the issue of math and science gender gaps, gender gaps in reading are in fact much larger and favor girls at both tails of the distribution. While gender gaps in reading literacy for the USA were smaller than those found internationally, the need for further progress remains. Enrolments of women outnumber men in college, with higher female GPA and completion rates than their male peers [16,63–66]. Raising the educational aspirations of boys who experience difficulties in reading literacy, and continuing support for early intervention is critical as a matter of gender equity.

**Mathematics Literacy**

Gender differences in mathematics literacy were comparatively larger for the United States than those found across other OECD nations. These findings are consistent with student test data reported by Hedges and Nowell [23], as well as findings from PISA 2003 [32,61] that a small gender difference in mathematics exists, but is also inconsistent with findings of no difference reported by Hyde and colleagues using data from the NAEP [20]. How are we to reconcile this discrepancy?

As reviewed earlier, problem-solving for complex and novel mathematics tasks show a small to medium sized male advantage [30], and PISA assessment of mathematical literacy is somewhat different to that of the NAEP. This may allow for greater differentiation between high and low ability students if a ceiling-effect is present, and may provide a more thorough test of the gender similarities hypothesis. It may well be the case that gender differences in basic mathematical literacy are trivial in size [71], but that gender differences can be found in more complex tasks [30] requiring more than just curriculum knowledge.

Gender differences were observed for US performance, $d = .22$, which is small in size by Cohen’s [53] conventions and non-trivial by Hyde’s [47] criteria. When expressed in the BESD format, the likelihood of being average or higher in mathematics increases from 44.5% for girls to 55.5% for boys. One should be careful not to make too much, or too little, of this gender difference. As Hyde [47] points out, the degree of overlap between male and female performance is large for effect sizes in the small range, with many girls performing at or above the male average in mathematics. This perspective does not diminish the observation that a gender

![Figure 2. Histogram of gender difference effect sizes in mathematics literacy across OECD nations.](doi:10.1371/journal.pone.0039904.g002)
gap exists. As can be seen from the cross-cultural evaluation of mathematics, gender gaps in mathematics are not an inevitability, with many countries in fact showing higher female performance.

This difference is most apparent when examining student attainment of the highest proficiency level in mathematics, with double the amount of boys than girls reaching this stage. Benbow [22] argued that gender differences in high-achievement for mathematics could be at least partially explained by greater male variability and a combination of biological and environmental factors. It is likely that greater male variability explains at least part of the gender difference in high achievement, but that sociocultural factors also play a role in the development of mathematics at the extreme tails of the distribution. While general proficiency in mathematics is an important life goal for all students, attainment of an advanced level of mathematics is an important prerequisite for pursuing more technical degrees in STEM-related fields [72]. A growing body of research suggests that self-efficacy and confidence in mathematics play an important part in the decision making process of women to pursue STEM-related careers or direct their talents elsewhere [23,62–64]. Increasing self-confidence in mathematics and instilling a sense of mastery may be a crucial component any educational intervention, as well as challenging negative cultural stereotypes about women’s ability in mathematics [41,69]. At least for students within the USA, gender differences in mean and high achievement for mathematics have not been eliminated, and highlight the need for further progress.

While cross-culturally, gender differences favored males across OECD and partner nations, the magnitude of this difference (d = .13) was also small in size and subject to wide cultural variation. The likelihood of being average or higher in mathematical ability increases from 46.7% for girls to 53.2% for boys, a small but non-trivial difference. Unlike reading

![Figure 3. Relationship between women in research and gender ratios of high-achievers in mathematics literacy.](doi:10.1371/journal.pone.0039904.g003)

**Table 7.** US National Science performance for girls and boys, including high and low achievers.

<table>
<thead>
<tr>
<th></th>
<th>Girls</th>
<th>Boys</th>
<th>Standard Deviation</th>
<th>Effect Size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>495</td>
<td>509</td>
<td>(98)</td>
<td>.14</td>
</tr>
<tr>
<td>OECD Average</td>
<td>501</td>
<td>501</td>
<td>(94)</td>
<td>.00</td>
</tr>
<tr>
<td>% students at lowest ability level, USA</td>
<td>4.6%</td>
<td>3.8%</td>
<td></td>
<td>1.20 girls : 1 boy</td>
</tr>
<tr>
<td>% at highest ability level, USA</td>
<td>1.0%</td>
<td>1.5%</td>
<td></td>
<td>1.52 boys : 1 girl</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0039904.t007
literacy, there were a number of countries which had non-significant gender differences, which would be inconsistent with strong biological differences between boys and girls in mathematical reasoning [11,15,63–66]. It may be the case that whatever slight advantage boys have is magnified by social and cultural reinforcement, to produce gender differences in some countries but that other nations raise girls and boys to equivalent performance.

A parallel may be also drawn between cross-cultural support for gender differences in mathematics, and similar evidence for gender differences in spatial ability [67–70]. Many theorists have argued that spatial ability provides a foundation for later development of mathematical ability [13,73–76]. Although gender differences are consistently found across all cultures favoring males, the magnitude of spatial differences is subject to cultural variation. In particular, Lippa, Collaer and Peters [77] compared national measures of gender equality and economic development with gender differences in spatial performance for a fifty-three nation sample, finding strong positive correlations with both measures. These findings are correlational, not causal, but taken together may change the way in which we think about the development of cognitive differences. It would appear
that gender differences in number of cognitive abilities are at least partially influenced by social and cultural influences such as gender equality and the status of women [32,61]. While parental, teacher and peer influences also play a part [78–83], the influence of wider cultural influences at the macro-level may be important considerations for any biopsychosocial models of gender difference.

Science Literacy

While the effect size for gender differences in science literacy for the USA was relatively small compared to that of reading and mathematics, it stands out as the largest effect size across all OECD nations, $d = .14$. This is a small effect size, but also not a trivial one by Hyde’s [47] conventions. Represented in the BESD format, the likelihood of being average or higher in science literacy increases from 46.5% for girls to 53.5% for boys. Additionally, boys were slightly overrepresented in attaining the highest level of science proficiency, but not to the same degree as for mathematics. Of all the domains assessed, science literacy appears to be the most variable cross-culturally, with many countries showing no difference whatsoever, and many showing a female advantage. This is a promising sign, and a benchmark to which the USA can aspire.

This pattern of results was consistent with the gender similarities hypothesis.

Gender Stratification Hypothesis

In order to test the gender stratification hypothesis, this study examined the relationship between national measures of gender equity and gender gaps in reading, mathematics and science literacy. While some support for the gender stratification hypothesis was found, the predictive validity of gender equity measures varied across instruments and domains. In particular, relationships between the Gender Gap Index instrument were often weak, and in the case of science literacy high achievers in a direction contrary to hypotheses. This failure to support the gender stratification hypothesis using all gender equity measures should not be interpreted as a refutation of the hypothesis, but means that one should evaluate the hypothesis carefully. Each instrument taps different aspects of the underlying gender equity construct, and it is likely that some elements of equity have greater bearing on educational outcomes than others. A consistent finding across all three domains, and across both mean performance and high achievers, was that the relative share of women in research accurately predicted the presence or absence of gender differences.
However, composite measures of gender equity showed weaker or inconsistent associations.

It may be the case that measures more closely related to education, such as gender differences in relative share of research and science positions, may more accurately measure the underlying social and cultural conditions that foster or inhibit the development of gender differences in reading, mathematics and science literacy. None of the instruments directly measure attitudes towards women in STEM-related fields, or gender stereotypes about the relative abilities of males and females [69,84]. Instead, the composite measures relate to the role of women in society in general, which may lack the specificity required to consistently predict gender differences in learning outcomes. Although increased gender equity generally may be associated with the presence or absence of gender gaps in reading, mathematics and science, it may not be the direct cause.

The relative share of women employed in scientific research may be more directly related to societal attitudes about the role of women in technical fields, and to gender stereotypes about the capabilities of males and females in sex-typed achievement domains. Girls growing up in a society that praises the scientific and technical achievements of men but lacks equivalent female role models may perceive that women are less capable in this area, or that their skills are not culturally valued. They may instead be motivated to develop other talents, such as high proficiency in language, and to pursue careers in less-segregated professions. Conversely, if girls grow up in a social environment where they see progression into further education and specialisation in STEM-related fields is not only possible but also commonplace, they may be more motivated to acquire and master mathematics and science skills. In such a culture, encouragement from parents and teachers may be higher, and they may show greater confidence and improved self-efficacy in these domains than children from other cultures. While mean gender differences are smaller (or favor females) in such nations, this also translates to increased female representation in high achievers as well. This provides for stronger support of the gender stratification hypothesis.

Economic Prosperity

Mean gender differences were larger for mathematics and science in economically prosperous nations as hypothesised but were largely unrelated to reading literacy. This likely reflects both increased educational spending for economically prosperous nations, as well as increased emphasis being placed on mathematics and science skills. Student achievement in less prosperous nations may be more homogenous with smaller gender differences, and there may be a reduced focus on teaching of these skills. It may also be the case that there is greater competition by males to achieve in these masculine sex-typed domains. These associations were also found for gender ratios in high achievement. Additionally, gender ratios for high achievers in reading literacy were also related to economic prosperity, which was unexpected.

Power Distance

Hofstede [50] argued that cultures differed in their tolerance for inequality, with some cultures observing social class distinctions more strongly than others. Such cultures may place greater emphasis on social roles and stratification, but one way of overcoming inequality is the pursuit of culturally valued skills and traits. As a compensatory strategy, girls may seek out higher social status positions by obtaining education in mathematics and science, and this may help to explain the female advantage for science observed for non-OECD nations. As hypothesised, these associations were found for mean gender differences in mathematics and science as well as for gender ratios of high achievers. Lesser support was found for cultural mediation in reading literacy, with no association for mean gender differences but a positive association for gender ratios in high achievement.

Social Implications

The question of whether gender differences exist in cognitive abilities has important implications for parents, educators, and policy-makers [20,47,72,82–83]. Yet great caution must be taken when interpreting empirical evidence - Hyde [47] raises a legitimate concern that inflated claims of wide gender difference might contribute to increased gender segregation in education and the workforce, and that the potential of girls may be overlooked by parents and teachers [78–92]. This study finds evidence of gender similarities rather than differences cross-culturally but also that meaningful gender gaps in maths and science remain and are related to cultural factors.

Society as a whole also has a vested interest in this question, both directly and indirectly. We as citizens rely on the services and advancements that a highly skilled science and technology workforce provide, with direct benefits for our health and lifestyle, and for an economy that depends on the brightest and most innovative of minds entering these fields to sustain an internationally competitive advantage. There are also indirect benefits from having a society that is at least partially scientifically literate – making decisions through the political process and personal choices about issues such as the use of stem-cell technologies, vaccination of children against disease, or evidence of climate change. When students, particularly girls, disengage with science learning there are costs to the individual, in the form of reduced security and income, but also to the wider society. While not every child may have the ability or interest to pursue a scientific career, a basic scientific literacy is required for full participation in society.

The underrepresentation of women in science is a serious social issue, and considerable resources are being expended to address this problem [72,83–84]. Recognising that a gender gap exists is the first step towards changing it, while cross-cultural evidence of gender similarities provides strong evidence that the gender gaps in maths and science are not inevitable. STEM-related careers can be a pathway to a higher standard of living and job security, and girls deserve the same encouragement as boys to pursue these professions as a matter of social justice. Newcombe et al. [85] argues that psychology can make a positive contribution to changing the social and educational environments that curtail the potential of all students in mathematics and science.

Strengths and Limitations

The broader coverage of nations included in the PISA 2009 round of assessment makes for a stronger test of research hypotheses than was previously possible. Additionally, many of the partner nations would be categorised as lower in human development, with reduced access to the educational advantages found in other nations. While researching educational outcomes for large and economically prosperous nations like the United States is important, debate about gender differences is often shaped by evidence from relatively affluent samples. In less advantaged nations, provided girls and boys are still afforded the same access to education, performance in maths and science literacy is more homogenous giving greater support to the gender similarities hypothesis. However, there is still substantial cultural variability in gender differences, and much of this is driven by cultural variation in gender equality. For a large portion of the world, the strongest predictor of gender differences in educational outcomes is equivalent access to education, occupational segrega-
tion, and representation of women in technical and research professions. If priority were to be given to improving these globally, substantial improvements in female literacy in maths and science could be realised.

While support for research hypotheses were generally observed, availability of data for cross-cultural correlations meant reduced statistical power to detect relatively weak correlations. It may well be the case that the hypothesised associations with mean gender differences across reading, maths, and science could have been detected with expanded coverage of Hofstede’s cultural dimensions [50]. There are likely many other cross-cultural correlates of gender differences that remain unexplored, such as gender stereotypes about cognitive abilities, and cultural variations in attitudes towards women in society. Such research is limited by the need to obtain wide coverage of these constructs across nations.

Summary
Evidence from national testing for the United States shows that there are meaningful gender gaps to be addressed in academic achievement across reading, mathematical and science literacy. Furthermore, these are larger than that found cross-culturally, where evidence for the gender similarities hypothesis is stronger. Globally, there is a small gender difference in mathematics literacy favoring males, and a small difference in science literacy favoring girls in non-OECD nations. However, a consistent finding for reading literacy is that girls outperform boys in both mean differences overall and gender ratios in attaining high reading achievement. Correlational analyses show that economic prosperity, gender equity, and the dimension of power distance are good predictors of global gender differences in cognitive abilities.

Author Contributions
Analyzed the data: DR. Wrote the paper: DR.


47. Sage Publications. 


Chapter 7 – Meta-Analysis of Sex-Role Mediation Effect for Visual-Spatial Ability

This study reports a meta-analysis of the sex-role mediation effect for visual-spatial ability. This chapter includes a co-authored paper that has been published as:


Permission for inclusion of the final paper has been granted by the publisher, Springer.

In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
- Data collection from archival sources
- Statistical Analysis
- Writing chapter

(Signed) ________________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ________________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann
Gender-Role Differences in Spatial Ability: A Meta-Analytic Review

David Reilly & David L. Neumann
Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media New York. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your work, please use the accepted author’s version for posting to your own website or your institution’s repository. You may further deposit the accepted author’s version on a funder’s repository at a funder’s request, provided it is not made publicly available until 12 months after publication.
Gender-Role Differences in Spatial Ability: A Meta-Analytic Review

David Reilly · David L. Neumann

Abstract Although gender-related differences in highly gender-typed cognitive abilities are of considerable interest to educators and cognitive researchers alike, relatively little progress has been made in understanding the psychological processes that lead to them. Nash (1979) proposed a gender-role mediation hypothesis for such differences, with particular emphasis on spatial ability. However, changes in gender equality and gender stereotypes in the decades since merit a re-examination of whether a gender-role association still holds (Feingold 1988). A meta-analysis of 12 studies that examined gender-role identity and mental rotation performance was conducted. These included studies from the United Kingdom, Canada, Poland, Croatia, and the United States of America. The mean effect size for masculinity was $r = .30$ for men and $r = .23$ for women; no association was found between femininity and mental rotation. This effect size was slightly larger than that found previously by Signorella and Jamison (1986), and exceeds many other factors known to influence spatial ability. The implications of gender-role mediation of gender differences are discussed and future research directions are identified.

Keywords Gender differences · Spatial ability · Gender-role mediation · Gender roles · Mental rotation · Meta-analysis

Introduction

Though progress has been made in closing gaps in recent decades, women still remain underrepresented in science, technology, engineering and mathematics (STEM)-related fields in the United States with fewer women entering these fields in tertiary education (National Science Foundation 2011). Concerns about the underrepresentation of women are also present in many other countries, including Britain (Brosnan 1998) and Australia (Bell 2010). Although exceptions exist for psychology and medical sciences (Hyde 2007b), in general women are underrepresented in the sciences at a graduate level, as well scoring lower in tests of mathematics and science achievement at school within the U.S. (Gallagher and Kaufman 2005; Hedges and Nowell 1995). These findings are also supported by more recent reviews of mathematics and science literacy in large international assessments of student achievement such as the Programme for International Student Achievement (Else-Quest et al. 2010; Guiso et al. 2008; Reilly 2012), which assesses students worldwide as they reach the end of compulsory schooling. Much of the research in this area, however, draws on samples from America, and all studies cited herein are U.S.-based unless otherwise noted.

A consensus statement issued by major researchers in the area of gender-related cognitive differences identified research into the sources of individual differences in STEM achievement as an important priority (Halpern et al. 2007). When men and women are compared at the population level, reviews find no evidence of gender differences in general intelligence (Halpern and Lamay 2000; Neisser et al. 1996). However, researchers have frequently observed gender differences in more specific components of cognitive ability (Boyle et al. 2010a, b; Neumann et al. 2007, 2010). The size of such differences ranges from small to large, as a function of the cognitive component under investigation (Halpern et al. 2011). The largest and most consistent gender differences are found in spatial ability (Halpern 2011; Kimura
Spatial Ability and Quantitative Skills

Many researchers (e.g., Wai et al. 2009) have proposed that spatial ability provides a foundation for the development of quantitative reasoning such as science and mathematics (Nuttall et al. 2005; Serbin et al. 1990). Factor analyses of cognitive ability tests show high loadings for mathematical performance against a spatial factor (Carrol 1993; Halpern 2000). Furthermore, measures of spatial ability have predictive validity, in that they can predict future performance in quantitative fields (Williams and Ceci 2007). For example, Shea et al. (2001) followed a large group of intellectually talented boys and girls over a 20-year longitudinal study, from seventh grade until age 33. They found that individual differences in spatial, verbal, and quantitative reasoning in adolescence predicted educational and vocational outcomes two decades later. Further, spatial ability made a significant unique contribution even after controlling for verbal and mathematical ability (Shea, et al. 2001). Spatial ability is also predictive of college mathematical entrance scores (Casey et al. 1995, 1997), which are an important prerequisite for entry to further education in science and mathematics disciplines (Ceci et al. 2009).

Factors that influence spatial ability during development hold promise for educational interventions that seek to reduce the gender gap in science and mathematics in adulthood (Halpern 2007; Newcombe 2007). Hyde and Lindberg (2007, p. 29) argued that even mild improvement in spatial ability may have “multiplier effects in girls’ mathematical and science performance”. Additionally, higher levels of spatial ability are associated with attitudinal changes towards mathematics and self-confidence in mathematical ability from elementary school (Eccles et al. 1993) to high school and college (Eccles 1987; Eccles et al. 1990). Thus the contribution of spatial ability to later cognitive development may be in part social as well as intellectual (Crawford et al. 1995; Nash 1979). Academic domains where one feels competent and are seen as being socioculturally valued for one’s gender are more likely to be pursued than those that are not (Eccles et al. 1990).

Although medium to large gender differences in spatial ability performance are found in most reviews of studies (Linn and Petersen 1985; Voyer et al. 1995), Hyde (2005) notes that within-gender variation is larger than between-gender differences. Since gender alone explains only a portion of individual variation in spatial ability (Caplan and Caplan 1994), identifying other developmental factors which promote spatial ability is an important research goal (Halpern et al. 2007; Hyde and Lindberg 2007). Neisser et al. (1996, p. 97) argued that understanding the source of such differences is critical, and that such questions are “socially, as well as scientifically important”. One potential source of individual differences is that of gender-role identity.

Although the exact mechanisms contributing to the emergence of gender differences in spatial ability are debated (see Caplan and Caplan 1994 and Halpern 2011 for a discussion) they are believed to be influenced by a network of biological and sociocultural contributions (Ceci et al. 2009; Crawford et al. 1995; Eagly and Wood 1999; Halpern and Tan 2001). One such contribution is that of gender-role identity.

Though boys and girls typically differ in early socialisation experiences (Eccles et al. 1990; Emmott 1985; Lytton and Romney 1991), there is considerable individual variation in the degree to which they develop and acquire stereotypically masculine and feminine personality traits, behaviors and interests (Bem 1974; Constantinople 1973; Kagan 1964a). This process is referred to as gender typing (Kohlberg 1966; Kohlberg and Ullian 1974), and holds implications for the development of gender-role identity and integration of masculinity and femininity into an individual’s self-concept and gender schema (Bem 1981; Knafo et al. 2005; Spence 1993). Highly gender typed individuals are motivated to keep their behavior and self-concept consistent with traditional gender norms (Bem 1975; Bem and Lenney 1976; Maccoby 1990; Martin and Ruble 2004), and this also applies to academic domains (Nosek et al. 2002; Oswald 2008; Steffens and Jelenec 2011). Others may integrate aspects of both masculine and feminine identification into their self-schema, termed androgyny (Bem 1984; Spence 1984).

Gender-Role Mediation of Spatial Ability

Nash (1979) proposed a gender-role mediation explanation for gender differences in which it is argued that gender-role identity can either promote or inhibit optimum development of cognitive ability in highly gender-typed domains, such as spatial and verbal ability. Specifically, Nash (1979) theorized that masculine identification leads to cultivation of spatial, mathematical, and scientific skills, whereas feminine identification facilitates verbal and language abilities.
In a review of gender-role influences on cognitive ability, Nash (1979, p. 263) wrote “For some people, cultural myths are translated into personality beliefs which can affect cognitive functioning in gender-typed intellectual domains”. This argument was based on earlier work by Sherman (1967) into differential learning and practice experiences of boys and girls. In doing so, Nash extended Sherman’s theory by placing cognitive development of spatial ability in a social context, where gender-role identity encourages or discourages optimum development of spatial potential. Nash identified several mechanisms that contribute to spatial development, including gender typing of intellectual domains, gender-role conformity and self-efficacy beliefs.

Differential Spatial Experiences

Sherman (1967) hypothesized a causal explanation for the presence of gender differences in spatial ability, based on a child’s differential opportunities to develop and refine spatial skills through play and recreational activities. Boys and girls typically differ in their socialisation experiences, and are encouraged by parents to engage in either stereotypically masculine or feminine play appropriate to their gender (Eccles et al. 1990; Lytton and Romney 1991). However, play is also an opportunity for active engagement and cognitive development (Piaget 1968). Caplan and Caplan (1994) argued that traditionally “masculine” typed activities promote the development of spatial ability by encouraging the practice and application of spatial skills (Connor and Serbin 1977). In contrast, traditionally “feminine” activities do not require the use of spatial skills, but reinforce other socially valued skills (Lever 1976).

What distinguishes Sherman’s (1967) explanation from other explanations (such as Caplan and Caplan 1994) is that it focuses specifically on gender roles, rather than solely on biological gender, as explaining individual differences in spatial ability. Differential practice of skills promoting spatial development occur through gender typing of activities and interests (Serbin and Connor 1979; Serbin et al. 1990). Rather than assuming that the lives of boys and girls do not overlap, or that all boys engage in a high level of activity and receive equal opportunities to practise and develop spatial ability, it accounts for individual differences and gender typing. There is evidence to support this argument. Retrospective studies have shown that an association exists between spatial ability and activity preferences in young adult college-level samples (Baenninger and Newcombe 1989; Signorella et al. 1989).

Gender Typing of Intellectual Domains

Kagan (1964b) noted that objects in the everyday world, social activities, and even intellectual pursuits become gender typed as either masculine or feminine, based on shared consensual beliefs that emerge very early in childhood. For example, reading and language is regarded as being feminine (Dwyer 1973, 1974), whereas mathematics, science and technology are regarded as masculine (Li 1999; Nash 1975). Both at an implicit (Lane et al. 2012; Nosek et al. 2009; Steffens and Jelenec 2011) and an explicit level (Benbow 1988; Halpern and Tan 2001), cultural beliefs about specific cognitive tasks as being inherently masculine or feminine prevail - even for generations growing up with increased gender equality (Liben et al. 2002). Recently, Halpern et al. (2011) showed that lay beliefs about cognitive gender differences in student and community samples were firmly entrenched across both men and women. Although these stereotypes are not an accurate reflection of reality, Nash (1979) argued they have the potential to shape the self-concepts of boys and girls, and how they see themselves in relation to these academic domains (Hyde and Lindberg 2007).

Gender-role Conformity Pressures

Gender roles and associated stereotypes describe differences between men and women, and prescribe how they should behave in social and occupational settings (Eagly and Mitchell 2004). Highly gender typed persons are motivated to keep their behavior consistent with internalised gender-role standards and norms (Bem and Lenney 1976), whereas those low in gender typing or for whom gender-role identity is less salient show greater cognitive and behavioral flexibility (Arbuthnot 1975; Bem 1975; Stein and Bailey 1973). Conformity cues as to who should engage in certain behaviors, and what activities are permissible for boys or girls, come from peers, parents, and the media (Martin and Ruble 2004; Matthews 2007), and this has implications for intellectual domains that are masculine or feminine dominated (Eccles 2007).

Nash (1979) argued that the increased saliency of gender and gender typing of academic subjects in adolescence may lead to a conflict between the “ideal” image a student holds of himself or herself, and the activities he or she chooses to perform well in and values. Perceived incompatibility between being “feminine” and succeeding in stereotypically “masculine” domains can hinder academic achievement (Rosenthal et al. 2011; Schmader 2002). Thus there is also an attitudinal and motivational component to development of intellectual abilities (Nash 1979).

Self-efficacy Beliefs and Gender Stereotypes

During childhood when gender-role saliency is low, boys and girls show relatively little difference in intellectual abilities, and what differences exist often favors girls (Halpern 2000; Nash 1979). However, gender typing of intellectual pursuits quickly emerges in adolescence (Dwyer 1974; Kagan 1964b),
Evidence for a Spatial-Gender-Role Association

A prior meta-analysis by Signorella and Jamison (1986) found support for Nash’s hypothesis in spatial ability. However there have been major and potentially relevant changes in gender roles and stereotypes in the intervening decades (Auster and Ohm 2000; Hyde and Lindberg 2007) which Feingold (1988) has argued are responsible for declining gender differences in cognitive ability. This view is supported by Hyde (2005, 2006, 2007) and colleagues across a range of intellectual abilities (Hyde 2007a; Lindberg et al. 2010). These changes question the validity of Nash’s theory in contemporary society and whether such gender-role associations still exist today. For this reason, we aimed to conduct a meta-analysis of studies published since Signorella and Jamison’s (1986) review, to see whether the gender-role mediation hypothesis still holds. Although these studies are primarily based on research conducted in the USA, studies from other nations (e.g., Poland, Croatia, United Kingdom, Canada) are also examined for a broader test of Nash’s theory.

Meta-analysis provides researchers with a way to critically evaluate the cumulative evidence of empirical evidence (Rosenthal 1984), and the technique is becoming increasingly common in psychology (Hyde 1990; Rosenthal and DiMatteo 2001). Although individual studies taken in isolation might show that a relationship between factor X on ability Y may be present or absent, factors such as random sampling error and lack of statistical power may result in erroneously rejecting the null hypothesis (Type I error) or failing to detect an effect that is real (Type II error). The technique of meta-analysis allows one to draw firmer conclusions about the existence of an association (Rosenthal and DiMatteo 2001), as well to arrive at an estimate of its size that is more accurate and reliable than could be determined from a single empirical study.

A requirement of meta-analysis is that empirical studies measure a similar construct drawn from similar samples (R. Rosenthal 1984, 1995), and that there are a sufficient number of studies to make meaningful conclusions. Spatial ability is not a unitary construct; it encompasses at least three separate processes – spatial perception, visualisation, and mental rotation (Linn and Petersen 1985). Mental rotation is one of the most widely researched areas of cognitive gender differences (Halpern and Lamay 2000), due in part to the fact comparisons of men and women in mental rotation show the largest effect sizes of all spatial tasks (Voyer et al. 1995). Some researchers regard mental rotation to be a representation of general spatial reasoning (Casey et al. 1995; Halpern 2000; Vandenberg and Kuse 1978), and there is evidence that performance in mental rotation prospectively predicts later development of quantitative reasoning (Casey et al. 1997; Nuttall et al. 2005). Therefore this review is confined to studies that investigated performance in mental rotation tasks. In addition, gender differences are larger after late adolescence when gender roles become particularly salient (Nash 1979). There are also issues of reliability and validity when assessing gender roles in younger samples. For this reason, only studies using high school, college or young adult samples were considered for inclusion in the reported meta-analysis. Studies using younger samples, such as that by Titze et al. (2010), were not considered.

In sum, the present review involved a meta-analysis of studies that have investigated gender-role associations with mental rotation task performance. It was hypothesized (Hypothesis 1) that masculinity would be positively associated with greater mental rotation performance in men and women. The influence of femininity was also investigated as a research question. It was hypothesised (Hypothesis 2) that there would be a negative association between femininity and mental rotation performance for both genders. Since the magnitude of gender differences typically varies with the type and level of difficulty of mental rotation task (Voyer et al. 1995), we also examined the type of mental rotation instrument as a potential moderator. Similarly, because there have been debates over which measures of masculinity and femininity are the best predictor of behavior (Bem 1984; Spence and Buckner 2000), we examined the type of gender-role instrument as a potential moderating variable.

Method

Search Strategy

To access as many studies as possible, a number of search strategies were used. Firstly, a Web of Science citation search for articles citing either Nash (1979) or Signorella...
Gender-role identity was measured using a psychometrically valid and reliable gender-role instrument, such as the Bem Sex Role Inventory (BSRI; Bem 1974) or the Personal Attributes Questionnaire (PAQ; Spence et al. 1974). Participants sampled were either an adult or high school aged adolescent, from a non-clinical sample. Requests to authors (n = 5) for additional information were made where a masculinity and mental-rotation association was not explicitly tested or reported. Three studies could not be included due to insufficient information to determine an effect size (Evardone and Alexander 2009; Tuttle and Pillard 1991; Vonnahme 2005). One practice sometimes adopted is to consider all studies missing an effect size to have an association with an absolute value of zero, a practice that Rosenthal (1995) considers overly conservative and leads to inaccurate estimates. This practice was considered at length by Hedges and Becker (1986) who cautioned against missing value substitution. Accordingly, the decision was made to exclude these missing studies. Following the application of the selection and exclusion criteria, there were 12 available studies examining mental rotation and gender roles. However, it should be noted that the possibility of unpublished null studies (commonly termed the “file drawer problem”) is addressed using meta-analytic techniques that test for publication bias (Orwin 1983; Rosenthal 1979).

Sample Characteristics

The characteristics of all studies identified in the literature search are presented in Table 1. Several of the studies recruited participants from different countries, making for a broader test of Nash’s hypothesis than would be possible if analyzing only data from the USA. It should be noted that in most studies, samples were drawn almost exclusively from student subject pools, limiting generalizability somewhat to a young-adult, college-level educated sample.

Procedure

Comprehensive Meta Analysis (CMA) V2 software was used for the calculation of statistics (Borenstein and Rothstein 1999). A random-effects model was chosen (Borenstein et al. 2009) because spatial ability is subject to a large number of psychosocial moderators, and a variety of different gender-role instruments and mental rotation tasks were used over multiple decades. The random effects model gives slightly wider confidence intervals than a fixed-effects model (Field 2001; Rosenthal and DiMatteo 2001), but gives a more appropriate estimate of how much variability is present in empirical studies (Kelley and Kelley 2012).

The focus of the review was the relationship between gender-role identity and mental rotation, which can be represented by Pearson’s product moment correlation, r. Gender-role instruments offer separate masculinity and femininity scales, allowing us to consider the effect of masculinity independently of femininity, and to test both for a mental rotation association.

Where the direct product–moment correlation between gender-role masculinity scale and mental rotation was reported, this was used because it represents the direct association independent of a subject’s femininity scale. However, two studies reported only the mean values for masculine, feminine, and androgynous groups. Since androgyny represents a “special case”, and some theorists argue that such participants cannot be legitimately combined with either the masculine or feminine group (Taylor and Hall 1982), the androgynous participants were excluded as per Signorella and Jamison’s (1986) recommendation. Such an approach is the most conservative strategy available, and may lead to an underestimation of the true effect size in cases where androgynous participants (high masculinity, high femininity) score higher than their masculine or feminine counterparts (e.g. Hamilton 1995). By doing so, however, it affords a simple comparison between masculine and feminine participants only, allowing for the use of Cohen’s d and then conversion to r as the common effect size unit using the formula given by Rosenthal (1984). Several studies recruited male or female participants only, and in several cases examined only masculinity associations. Calculations were performed using the CMA software.
<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Sample Type</th>
<th>Sample Age</th>
<th>N</th>
<th>Gender Role</th>
<th>Mental Rotation</th>
<th>Males M (r)</th>
<th>Females F (r)</th>
<th>Males M (r)</th>
<th>Females F (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jamison and Signorella (1987)</td>
<td>USA</td>
<td>High school students</td>
<td>8th grade</td>
<td>10 females</td>
<td>BSRI</td>
<td>CRT</td>
<td>.45</td>
<td>.04</td>
<td>.27</td>
<td>.14</td>
</tr>
<tr>
<td>Signorella et al. (1989)</td>
<td>USA</td>
<td>Subject pool</td>
<td>n/r</td>
<td>132 females</td>
<td>BSRI</td>
<td>CRT</td>
<td>.08</td>
<td>.01</td>
<td>.24</td>
<td>.05</td>
</tr>
<tr>
<td>Gilger and Ho (1989)</td>
<td>USA</td>
<td>Subject pool</td>
<td>M=19.0</td>
<td>52 females</td>
<td>BSRI</td>
<td>TSRT</td>
<td>.00</td>
<td>-.17</td>
<td>.00</td>
<td>-.17</td>
</tr>
<tr>
<td>Voyser and Bryden (1990)</td>
<td>Canada</td>
<td>Subject pool</td>
<td>M=21.0</td>
<td>65 females</td>
<td>BSRI</td>
<td>VMRT</td>
<td>.50**</td>
<td>-</td>
<td>.21</td>
<td>-</td>
</tr>
<tr>
<td>Tuttle and Pillard (1991)</td>
<td>USA</td>
<td>Community</td>
<td>Range 25–40</td>
<td>88 females</td>
<td>CPI</td>
<td>TSRT</td>
<td>n/r</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newcombe and Dubas (1992)</td>
<td>USA</td>
<td>Longitudinal</td>
<td>16</td>
<td>61 females; attrition rate = 29 %</td>
<td>PAQ</td>
<td>TSRT</td>
<td>-</td>
<td>-</td>
<td>.15</td>
<td>-.10</td>
</tr>
<tr>
<td>Hamilton (1995)</td>
<td>United Kingdom</td>
<td>Community, school and college</td>
<td>M=18.0</td>
<td>122 females</td>
<td>BSRI</td>
<td>SMRT</td>
<td>.12</td>
<td>-.12</td>
<td>.14</td>
<td>-.14</td>
</tr>
<tr>
<td>Jagieka and Herman-Jeglinska (1998)</td>
<td>Poland</td>
<td>Subject pool</td>
<td>n/a</td>
<td>30 males</td>
<td>BSRI</td>
<td>SMRT</td>
<td>.34*</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Saucier et al. (2002)</td>
<td>Canada</td>
<td>Subject pool</td>
<td>M=22.8</td>
<td>54 females</td>
<td>PAQ</td>
<td>VMRT</td>
<td>.45**</td>
<td>-.02</td>
<td>.45***</td>
<td>-.02</td>
</tr>
<tr>
<td>Rahman et al. (2004)</td>
<td>United Kingdom</td>
<td>Community</td>
<td>Range 18–40</td>
<td>120 females</td>
<td>EPP</td>
<td>VMRT</td>
<td>.41***</td>
<td>-</td>
<td>.23***</td>
<td>-</td>
</tr>
<tr>
<td>Ritter (2004)</td>
<td>United Kingdom</td>
<td>Subject pool</td>
<td>M=21.0</td>
<td>37 females</td>
<td>BSRI</td>
<td>SMRT</td>
<td>.34*</td>
<td>-.26</td>
<td>-.18</td>
<td>-.14</td>
</tr>
<tr>
<td>Scarbrough and Johnston (2005)</td>
<td>USA</td>
<td>Subject pool</td>
<td>M=19.6</td>
<td>41 females</td>
<td>BSRI</td>
<td>CSMRT</td>
<td>-</td>
<td>-</td>
<td>.40**</td>
<td>.00</td>
</tr>
<tr>
<td>Vonnahme (2005)</td>
<td>USA</td>
<td>Subject pool</td>
<td>M=21.2</td>
<td>46 males</td>
<td>BSRI</td>
<td>CMRT</td>
<td>n/r</td>
<td>n/r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hromatko et al. (2008)</td>
<td>Croatia</td>
<td>Unspecified</td>
<td>M=24.8</td>
<td>26 females</td>
<td>BSRI</td>
<td>TSRT</td>
<td>-</td>
<td>-</td>
<td>.64***</td>
<td>.03</td>
</tr>
<tr>
<td>Evardone and Alexander (2009)</td>
<td>USA</td>
<td>Subject pool</td>
<td>M=20.0</td>
<td>52 females</td>
<td>BSRI</td>
<td>VMRT</td>
<td>n/r</td>
<td>n/r</td>
<td>n/r</td>
<td>n/r</td>
</tr>
</tbody>
</table>

*a* Calculated from p value  
*b* Androgynous (high masculinity, high femininity) eliminated * p < .05; ** p < .01; *** p < .001 two-tailed  
*c* Calculated from median split of masculinity/femininity  
*d* Data provided by author  
*e* Data only available for masculinity n/r = data not reported

*CPI* California Psychological Inventory; *EPP* Eysenck Personality Profile (EPP; Eysenck et al. 1996); *CRT* Card Rotation Test (French et al. 1963); *TSRT* Thurstone Spatial Relations Test (Thurstone 1958); *SMRT* Shepard and Metzler Mental Rotation Test (Shepard and Metzler 1971); *VMRT* Vandenberg Mental Rotation Test (Vandenberg and Kuse 1978); *CMRT* Cooper and Shepard MRT (Cooper and Shepard 1973)
**Meta-analytic Results**

Study characteristics and effect sizes are presented in Table 1. Since empirical studies using gender roles frequently find gender × gender-role interactions, the associations with masculinity and femininity are reported separately for men and women. Forest plots are provided when gender-role associations are statistically significant. A forest plot conveys a visual representation of the effect size estimates of individual studies and their variability (Lewis and Clarke2001); one can see the amount of variation between individual studies as well as the overall trend. In the centre of each study’s confidence interval is a square; the size of the square corresponds to the sample size used in each study. The diamond symbol represents the overall estimate of the sample, with the centre of the diamond being the point estimate and its horizontal tips representing the confidence interval.

**Girls and Women**

Figure 1 presents a forest plot of the association between masculinity and mental rotation performance for girls and women, and effect sizes are given in Table 1. Hypothesis 1 predicted that masculinity would be positively associated with greater mental rotation performance. As shown in Fig. 1, most studies with female samples were in a direction consistent with this hypothesis with the exception of two studies: Gilger and Ho (1989) found no association, whereas Ritter (2004) found a weak negative association. The distribution of effect sizes across studies was heterogenous, $Q(10) = 21.13, p = .020, I^2 = 52.67$ indicating moderate variability across studies. It is also noteworthy that the two largest associations were found in the non-USA samples of Croatia ($r = .64$) and Canada ($r = .45$). However, the size of the correlation is unlikely to be culturally related given that the third largest association was found in a USA sample ($r = .40$) and that small associations were also found in non-USA samples (e.g., $r = -.18$ for Ritter 2004; $r = 14$ for Hamilton 1995).

In support of Hypothesis 1, the combined masculinity effect size for women was $r = .23$ (95 % CI lower = .11, upper = .34), $Z = 3.72, p < .001$. This correlation for women was only slightly larger than that found by Signorella and Jamison (1986), who found a significant association of $r = .19$ between masculinity and mental rotation for girls and women using androgyny measures. To put these findings into perspective, we employed Rosenthal’s Binomial Effect Size Difference (BESD; R. Rosenthal and Rubin 1982), a metric that represents effect size in a format suitable for interpretation by non-statisticians (R. Rosenthal and DiMatteo 2001). Represented in the BESD format, the likelihood of being average or higher in mental rotation performance increases from 38.5 % for feminine women to 61.5 % for masculine or androgynous women.

The possibility of unpublished null studies (referred to as the “file drawer problem”) was also addressed by the calculation of Orwin’s Fail-Safe $N$, which estimates the number of null studies required to reduce mean effect sizes to a specific cutoff-point (Borenstein et al. 2009; Orwin 1983). Employing Orwin’s calculation, it would take only two more null studies to reduce the association to that found previously by Signorella and Jamison (1986); therefore the stronger association in these studies should be taken only tentatively.

Hypothesis 2 predicted that there would be a significant negative association between femininity and mental rotation performance. This hypothesis was not supported, $r = -.05, p = n.s.$ Such a finding is also consistent with the findings of Signorella and Jamison (1986) who failed to find any association between femininity and mental rotation performance.

**Boys and Men**

The forest plot of the association between masculinity and mental rotation performance for boys and men is shown in Fig. 2 and it presents the second test of Hypothesis 1. As can
be seen from the figure, the scores of men were slightly wider in variability than for women, with many studies showing relatively large associations while three showed relatively weak or non-significant correlations. Similar to the results for women, there did not appear to be a strong relationship between the country the study was conducted in and the size of the association. The largest association was found in a sample of men from Canada ($r = .50$), but the equal second largest association was found in a USA male sample ($r = .45$). However, it is noteworthy that the two remaining studies with USA samples did not find any significant association ($r = .08$ and $r = .00$). The distribution of effect sizes across all studies was heterogenous, $Q(8) = 17.92, p = .022$, $I^2 = 55.36$, indicating moderate variability between studies.

In support of Hypothesis 1, the association between masculinity and mental-rotation performance for men was significant, $r = .30$, (95% CI lower = .16 upper = .42), $Z_{ma} = 4.25$, $p < .001$. Again, the association is slightly larger than that estimated by Signorella and Jamison (1986), who reported an $r = .15$ between masculinity and mental rotation performance for boys and men. Orwin’s Fail-safe $N$ showed that it would take an additional eight unpublished studies with a mean association of zero to reduce this correlation to the size found in the earlier review ($r = .15$). Represented in the BESD format, the likelihood of being average or higher in mental rotation performance increases from 35% for feminine boys and men to 65% for those with a masculine or androgynous gender-role identity. Finally, in contrast to Hypothesis 2, no association was found between femininity and mental rotation for boys and men, $r = -.06$, $p = n.s.$

**Moderating Variables**

Since there was moderate between-study heterogeneity in the masculinity association for both men and women, it is important to determine potential moderators that may be responsible such as the type of gender-role instrument used to classify participants, or the nature of the mental rotation task. Alternately, instruments might vary in their predictive validity for men and women, and this information might be useful in planning future research. Accordingly, effect sizes and heterogeneity were examined for men and women separately across gender-role instrument.

Tables 2 and 3 present associations across type of gender-role instrument for men and women respectively. While the BSRI was used most frequently, the strongest gender-role associations were found with the PAQ for both men and women. However with an insufficient number of studies employing gender-role measures other than the BSRI, any conclusions made about the predictive validity of these instruments are tentative.

Another potential source of heterogeneity is the nature of the mental rotation task employed. Meta-analytic reviews have found that the magnitude of gender differences differs across instruments (Voyer et al. 1995). It seems likely, therefore, that similar variation would be present when considering gender-role associations. Table 4 presents effect sizes for studies grouped by mental rotation instrument. Instruments were grouped into four categories. These groupings reduced heterogeneity, suggesting that much of the variability observed across studies was the result of using different instruments for measuring mental rotation. It should also be noted that the Vandenberg instrument also produced the highest gender-role effect size of any mental rotation task. This may reflect the increased difficulty of this instrument which allows for greater differentiation between high and low ability (Voyer et al. 1995).

**Discussion**

The present meta-analysis examined evidence for Nash’s (1979) gender-role mediation hypothesis of spatial ability, as measured by performance on mental-rotation tasks. In a previous review, Signorella and Jamison (1986) found a small but statistically significant association between gender role and mental rotation performance. The present results support the conclusions drawn by Signorella and Jamison (1986). There is a significant and medium sized association between masculinity and mental rotation in research conducted in the past 25 years. The size of the association did not appear to be strongly related to the country in which the study was conducted, although there was some evidence that the type of mental rotation task and gender-role measure used in the study was a factor. The present meta-analysis also showed that there was no association between femininity and mental rotation performance.

**Table 2** Effect size and heterogeneity by gender-role instrument for men

<table>
<thead>
<tr>
<th>Type of instrument</th>
<th>N of studies</th>
<th>Effect size ($r$)</th>
<th>$Z_{ma}$, $p$-value</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bem Sex-Role Inventory</td>
<td>7</td>
<td>.25</td>
<td>$Z = 3.07, p = .002$</td>
<td>$Q(6) = 11.73, p = n.s.$</td>
</tr>
<tr>
<td>Personal Attributes Questionnaire</td>
<td>1</td>
<td>.45</td>
<td>$Z = 2.20; p = .028$</td>
<td>N/A</td>
</tr>
<tr>
<td>Eysenck Personality Profiler</td>
<td>1</td>
<td>.41</td>
<td>$Z = 2.50; p = .013$</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Total heterogeneity within-groups, $Q(6) = 11.74, p = .068$; between-groups, $Q(2) = 6.18, p = .045$
The results of this meta-analysis demonstrate three important things. Firstly, it upholds the claims made by Nash (1979) that, at least for mental rotation tasks, masculine gender roles contribute to the development of spatial ability. Although only correlational in nature, the inclusion of the longitudinal study by Newcombe and Dubas (1992) shows that gender roles have predictive validity for later development of spatial ability. Secondly, this review demonstrates the persistence of gender roles over a larger span of time, in that studies reviewed are drawn from three decades of research; it would appear that the empirical findings of Nash and others were not a statistical quirk, or an artefact of prevailing gender inequalities of the past. Thirdly, the review shows that the magnitude of the gender-role association may be somewhat larger than previously thought by researchers, especially for men.

A possible explanation for finding a stronger association between gender roles and spatial ability than Signorella and Jamison (1986) is the quality of instruments used across studies. Many of the earlier studies reviewed by Signorella and Jamison (1986) used instruments that operationalised masculinity and femininity as bipolar opposites of a unidimensional construct (Constantinople 1973) rather than orthogonal aspects of gender-role identity (Bem 1981). This leads to misclassification of masculine, feminine, and androgynous participants (Bem 1974, 1977) and an attenuation of effect size due to imprecision (Cooper 1981). It is difficult to suggest a theoretical reason why gender roles might influence cognitive development more strongly now in men than in previous decades, but this possibility cannot be ruled out entirely.

A growing trend in empirical research is a move away from levels of statistical significance towards evaluations of the magnitude of effect sizes (Wilkinson 1999), to assess their practical impact and importance. Cohen (1988) provides a good rule of thumb to gauge associations by: correlations of .10 or higher are regarded as small, .30 or higher as medium, and correlations higher than .50 are considered large. Frequently these yardsticks are used rather rigidly, and some researchers regard differences that are “small” as “trivial” or non-existent (Hyde 1996, 2005). Cooper (1981) warns against this practice, as the magnitude of effects that may be found can differ greatly from one field of psychological research to another. Similarly, in a review of effect sizes and practical importance for research with children, McCartney and Rosenthal (2000) advise against such yardsticks, and caution that effect sizes should be compared to those found in that particular research domain. For this reason, comparisons to a range of other effects deemed previously to be influential in spatial ability may be better able to put the results of this meta-analysis in context (Hyde 1990).

The present results showed a gender-role association of \( r = .30 \) for men and \( r = .23 \) for women. Two areas previously documented to contribute to spatial ability are prior spatial activity preferences in childhood (Signorella et al. 1989) and socioeconomic status (Levine et al. 2005). A meta-analysis by Baenninger and Newcombe (1989) produced an \( r = .10 \) between spatial activity preferences and spatial ability. Levine et al. (2005) found that spatial ability differences are found between low, medium, and high socioeconomic status groups for adolescents with an effect size \( r = .23 \) for mental rotation. When compared to these factors, which researchers have previously argued to be important and have a meaningful impact on spatial ability, the contribution of gender role and mental rotation is greater, and may go some
way to explaining existing gender differences in spatial ability (Nash 1979; Sherman 1967).

Implications for Spatial Development

One possible intervention being considered for individuals most at risk of forestalled spatial development is that of spatial training. In a review article, Newcombe and Frick (2010) stress the importance of early intervention in the development of spatial abilities during early childhood. Although it would be desirable to offer spatial instruction and training for all students to address this gender-gap (Hyde and Lindberg 2007), competing interests in an ever crowded curriculum make the likelihood of this practice being adopted rather bleak; indeed few schools incorporate spatial ability specifically into the curriculum during elementary school (Mathewson 1999; Newcombe and Frick 2010). A more practical measure might be for limited intervention programs to target at-risk students, in the same way that reading and literacy interventions are offered for students struggling in these areas.

While screening directly for spatial deficits may be possible, large gender differences do not typically emerge until adolescence (Linn and Petersen 1985; Voyer et al. 1995). Early intervention is desirable before such differences emerge (Newcombe and Frick 2010). The assessment of gender roles might serve as a more useful risk factor to consider than gender, and it has the advantage of not necessarily being limited to one gender. Nuttall et al. (2005) describe gender-role appropriate intervention programs that develop spatial expertise, but as of yet, there are no longitudinal studies of such programs. Educators may wish to be mindful to include a range of opportunities that encourage spatial development as well as stressing their importance and relevance to both boys and girls.

Newcombe and Frick (2010) also advocate early intervention by parents, in providing children with activities and opportunities outside the classroom to develop spatial awareness, perception and visualisation. Rigidly held gender roles restrict children’s self-selection of activities (Ruble et al. 2006; Tracy 1987), and parents may wish to encourage a broader repertoire in their children including sports and toys that encourage spatial development (Doyle et al. 2012). The continuing failure to find a negative relationship between femininity and spatial ability for both genders is also noteworthy. Feminine identification should not be discouraged in order to develop spatial and quantitative ability.

Future Directions for Research

Although gender differences in cognitive ability are frequently debated, many researchers note there is greater within-gender variability than between men and women (Hyde 1990; Priess and Hyde 2010). Gender-role identity appears to be an important, but previously underestimated contributor to these individual differences in spatial ability, which in turn is a key foundation for higher-level quantitative skills such as mathematics (Casey et al. 1997; Delgado and Prieto 2004) and STEM related fields (Halpern 2007; Newcombe 2007). Indeed, Halpern (2007, p. 125) has claimed that spatial ability is “essential” for success in STEM-related subjects. As such, the emergence of gender roles as a factor that meets or exceeds other factors that contribute to spatial ability is important, both as a potential diagnostic indicator for interventions as well as a focus for future investigation. By better understanding the psychosocial processes associated with gender roles and intellectual development, one might be able to identify strategies - such as self-efficacy training or challenging of gender stereotypes - that would help negate performance impairments.

Additionally, this meta-analysis affirms the merit of considering gender roles, rather than just biological gender, in studies of individual differences in cognition. Though this review was confined to only mental-rotation, it remains to be seen whether the results can be generalised more widely to other spatial ability tasks such as spatial perception and visualisation (Linn and Petersen 1985). For example, is there something specific about a masculine or androgynous gender role that leads to improved ability to perceive spatial objects and mentally rotate them, or can it be generalised to other spatial tasks? This would allow us to test whether gender-role differences in perception are chiefly responsible, or whether there are differences in the actual cognitive processes underlying such tasks, for example a general cognitive style (Arbuthnot 1975; Milton 1957). A limited number of studies with adolescents and young adults have considered the Piaget water-level task (Jamison and Signorella 1980; Kalichman 1989; Popiel and De Lisi 1984; Signorella and Jamison 1978) or the Embedded Figures Test (Bernard et al. 1990; Brosnan 1998; Hamilton 1995), with some inconsistencies, but larger studies are required. Furthermore, as Signorella and Jamison (1986) note, Nash’s (1979) hypothesised associations between gender-role identity and verbal ability remain largely untested, which future studies should pursue.

Conclusion

We have seen many changes in society’s beliefs about gender equality in the intervening decades since Nash (1979) proposed her gender-role mediation hypothesis of intellectual development. However, for spatial ability at least, this association seems as relevant today as when the claim was first made. The results from our meta-analysis support Nash’s hypothesis for the development of spatial ability, and this provides strong support for calls to conduct further research in this area to investigate the cognitive and social processes that underlie the association between gender-roles and cognitive abilities.
Acknowledgments This research was supported in part by a Griffith University Postgraduate Research Scholarship. Thanks go to Dr Heather Green, Dr Michael Steele, Dr Elizabeth Conlon, and Dr Margaret Signorella for early revisions of this manuscript.

References


Meta-Analysis Summary and Prelude to Empirical Studies

The previous set of chapters sought to examine whether the previously observed sex differences in specific cognitive abilities (verbal and quantitative reasoning) would still exist in contemporary samples, or alternatively, if they’d been eliminated as claimed by Feingold (1988), Hyde (2005), as well as Caplan and Caplan (1997, 2016). Additionally, it sought to contextualise that difference by evaluating the magnitude of observed sex differences, and determine if they were large enough to have practical importance. For quantitative reasoning, small but not trivial mean sex differences were found for mathematics and more substantial gender gaps in high achievers. Somewhat larger mean sex differences were found for science achievement, and again a sharp disparity in the tail ratios for high achievers. For verbal and language abilities, substantial sex differences were found for reading and writing with a developmental trend observed with age/years of schooling. Examination of tail ratios also showed substantial gender gaps in low- and high-achievers.

Collectively these studies provided a rationale for further investigation to evaluate support for the sex-role mediation hypothesis. Signorella and Jamison (1986) had conducted a meta-analysis on the association between masculinity and visual-spatial ability, but the literature was now dated. For this reason, we produced a meta-analysis of the association between masculine sex-role identification and visual-spatial ability with more recently collected data. Additionally, this was useful in contextualising the expected effect size for statistical power calculations in the empirical study that follows.
Chapter 8 – Empirical Study 1 – Sex and Sex-Role Differences in Specific Cognitive Abilities

“If women are expected to do the same work as men, we must teach them the same things.” – Plato, The Republic.

This study reports the empirical study into sex and sex-role differences in verbal and visual-spatial abilities. Specifically it tests Nash’s (1979) sex-role mediation hypothesis in a modern sample, finding support for both predicted tranches (verbal and visual-spatial ability) across a range of tasks. This has been published as:


Permission for inclusion of the final paper has been granted by the publisher, Elsevier. In accordance with the Griffith University Code for the Responsible Conduct of Research, a statement of contribution is provided for authorship of this paper. I acknowledge the contribution of my supervisors to this manuscript.

My contribution involved:
- Data collection from archival sources
- Statistical Analysis
- Writing chapter

(Signed) __________________________________________ (Date) : 1/12/18
David Reilly

(Countersigned) ________________________________ (Date) : 1/12/18
Primary Supervisor David L. Neumann

(Countersigned) ________________________________ (Date) : 1/12/18
Associate Supervisor Glenda Andrews
Sex and sex-role differences in specific cognitive abilities

David Reilly a,⁎, David L. Neumann a, Glenda Andrews a,b

a School of Applied Psychology, Griffith University, Queensland, Australia
b Menzies Health Institute Queensland, Queensland, Australia

Abstract

Sex differences in cognitive abilities are a controversial but actively researched topic. The present study examined whether sex-role identity mediates the relationship between sex and sex-typed cognitive abilities. Three hundred nine participants (105 males and 204 females) were tested on a range of visuospatial and language tasks under laboratory conditions. Participants also completed measures of sex-role identity, used to classify them into masculine, feminine, androgynous and undifferentiated groups. While sex differences were found for some but not all measures, significant sex-role differences were found for all spatial and language measures with the exception of a novel 2D Mental Rotation Task. Masculine sex-roles partially mediated the relationship between sex and a composite measure of spatial ability, while feminine sex-roles fully mediated the relationship between sex and a composite measure of language ability. These results suggest that sex-role identity may have greater utility in explaining individual differences in cognitive performance than biological sex alone.

Introduction

The topic of sex differences in cognitive abilities remains an active but controversial research question because of its educational, social and policy implications (Eagly & Mitchell, 2004; Halpern, 2014). While most reviews find that males and females do not differ in general intelligence (Halpern, Beninger, & Straight, 2011; Jensen, 1998; cf. Nyborg, 2015) sex differences are frequently found in specific cognitive abilities (Nisbett et al., 2012). Robust and sizeable sex differences are found for visuospatial ability (referred herein as spatial ability) and verbal ability (Miller & Halpern, 2013). Overall, males do better on spatial tasks such as mental rotation and spatial perception (Voyer, Voyer, & Bryden, 1995), while females do better on language tasks such as verbal fluency and grammar (Halpern & Lamay, 2000; Lynn, 1992). The effect sizes are moderately large, and are reflected in beliefs about gender differences in cognitive ability (Halpern, Straight, & Stephenson, 2011). Spatial and verbal skills are of particular interest to educational researchers for two reasons. Firstly, research suggests that spatial ability forms the basis for the development of sex differences in quantitative reasoning such as mathematics and science (Newcombe & Frick, 2010; Wai, Lubinski, & Benbow, 2009). Despite significant progress in closing the gender gap, meaningful sex differences in mathematics and science achievement persist, at least for students in the USA (McGraw, Lubinski, & Strutchens, 2006; Reilly, Neumann, & Andrews, 2015). This is an active area of research, given the underrepresentation of women in science, technology, engineering and mathematics (collectively referred to as STEM) fields (National Science Foundation, 2011). Furthermore, international assessments of student achievement such as the OECD’s Programme for International Student Assessment (PISA) also find sex differences in mathematics and science for some, but not for all, nations (Else-Quest, Hyde, & Linn, 2010; Guiso, Monte, Sapienza, & Zingales, 2008; Reilly, 2012). Secondly, verbal ability and language competence are essential life skills required for full participation in society and the workforce. Both within the United States, and cross-culturally, males consistently score significantly lower than females on tests of reading and writing (Guiso et al., 2008; Klecker, 2006; Lynn & Mikko, 2009; Reilly, 2012). Some researchers have speculated that this contributes to the growing trend across most Western nations of fewer men than women entering and completing tertiary education (Alon & Gelbgiser, 2011; Buchmann & DiPrete, 2006). Thirdly, both spatial and verbal abilities are specific cognitive abilities that are frequently investigated by sex researchers, and emerge as distinct separate factors of intelligence (Johnson & Bouchard, 2007).

1. Theoretical perspectives on sex-typed cognitive abilities

When sex differences are observed by researchers, this raises questions regarding their origins (Wood & Eagly, 2000). Early research into sex differences in cognitive abilities focused primarily on biologically-based explanations, including the contribution of hormones (Auyeung et al., 2009; Hines, 1990; Kimura & Hampson, 1994) and anatomical structures such as the corpus callosum (Hines, Chiu, McAdams, Bentler, & Lipcamon, 1992). One argument supporting such a view is the observation of greater male variability (Feingold, 1992; Machin & Pekkinen, 2008), leading to exaggerated sex differences at the extreme tails of the ability distribution. While sex differences in the extremely gifted is an important topic in its own right, as they

Keywords:
Sex differences
Sex-role mediation hypothesis
Spatial ability
Verbal ability
related to a relatively small percentage of the population, the majority of sex difference research is concerned with mean sex differences between males and females as a group. Empirical studies into the effect of hormones on cognition find mixed support (cf. Halari et al., 2005; Kimura & Hampson, 1994), and that they explain only a small percentage of variance. In recent decades explanations have broadened to incorporate sociocultural factors, such as differences between boys’ and girls’ early socialization experiences (Lyttton & Romney, 1991), differential parental expectations for sons and daughters (Eccles, Jacobs, & Harold, 1990; Furnham, Reeves, & Budhani, 2002), gender stereotypes (Archer, 1992; Shapiro & Williams, 2012), and cultural beliefs (Guiso et al., 2008; Reilly, 2012). Most researchers now accept that sex differences are influenced by a network of biological and sociocultural factors rather than any single factor (Ceci, Williams, & Barnett, 2009; Nisbett et al., 2012; Wood & Eagly, 2012).

2. Sex role mediation of cognitive abilities

While it is difficult to disentangle nature from nurture, a commonality that is shared by both is that they contribute towards the development of an individual’s sex-role identity or the degree to which an individual embodies stereotypically masculine and feminine personality traits, behaviors, and interests (Bem, 1981b; Spence & Buckner, 2000). Though boys and girls as two distinct groups will differ in their early socialization experiences (Lyttton & Romney, 1991; Martin & Ruble, 2004), there is considerable individual variation within each gender group in the degree to which a person acquires sex-typed traits. While some children become rigidly sex-typed, others incorporate elements of both masculinity and femininity into their persona (Wood & Eagly, 2015). Highly sex-typed individuals are motivated to keep their behavior and self-concept consistent with traditional gender norms (Bem & Lenney, 1976; Martin & Ruble, 2004), including the sex-typing of specific skills, interests, and cognitive abilities.

Nash (1979) proposed the sex-role mediation hypothesis as one such explanation for the origins of sex differences in specific cognitive abilities. Nash (1979, p. 263) wrote “For some people, cultural myths are translated into personality beliefs which can affect cognitive functioning in sex-typed intellectual domains”. This argument was based on earlier work by Sherman (1967) into differential learning and practice experiences of boys and girls. Under the sex-role mediation hypothesis, masculine identification promotes the development of spatial reasoning and mathematics, while feminine identification promotes verbal ability and language aptitude (see Fig. 1). Essentially, the sex-role mediation hypothesis proposes that group differences in cognitive abilities emerge as a result of individual differences in sex-role identification (Durkin, 1987).

There is evidence to support sex-role mediation, at least for the development of spatial ability. Reilly and Neumann (2013) conducted a meta-analysis of the association between masculinity and mental rotation (the most commonly used measure of spatial ability), finding a robust association for both males and females. However, it is unclear whether such an association generalizes to other types of spatial ability such as spatial perception and visualization. An earlier review by Signorella and Jamison (1986) found an association with these types of measures, but it is unclear whether a similar result would be found in modern samples. Furthermore, few studies have investigated the second aspect of Nash’s sex-role mediation hypothesis, namely that feminine identification promotes the development of reading and language skills. Indeed, Signorella and Jamison noted that there was “a paucity of studies” (p. 219) that provide a test of sex-role mediation with language tasks.

3. The present study

The aim of the present study is to test the sex-role mediation hypothesis across a broader range of spatial and verbal tasks than previously used by researchers. There have also been considerable changes in the roles of men and women with the passage of time, so it is arguable whether historical conceptualizations of masculinity and femininity still apply (Auster & Ohm, 2000; Hoffman & Borders, 2001). Furthermore, some researchers have claimed that the magnitude of sex differences is diminishing in response to these social changes (Priess & Hyde, 2010). However, implicit gender stereotypes about sex-typing of cognitive tasks as being either masculine or feminine remain strong (Martin & Ruble, 2004; Nosek, Banaji, & Greenwald, 2002), as do beliefs about cognitive sex differences (Halpern, Straight, et al., 2011). We set out to determine whether previous experimental studies finding evidence of sex-role mediation (e.g. Hamilton, 1995) would be replicated when recruiting from a modern sample of young adults.

Linn and Petersen (1985) categorized tests of spatial ability as falling into one of three domains: mental rotation, spatial perception, and spatial visualization. The largest sex differences are found in mental rotation, while spatial perception also shows appreciably large sex differences (Voyer et al., 1995). However, the skill of spatial visualization shows relatively small sex differences which are sometimes not statistically significant, and so is less seldom included in a battery of cognitive measures. We selected measures from all three spatial domains (rotation, perception and visualization) so as to provide good content validity of spatial reasoning. We also employed a second test of mental rotation using two dimensional objects as stimuli, as most mental rotation tasks employ three dimensional objects at a cost of increased task difficulty.

The range of tasks available for measuring verbal ability is broad and less neatly defined than for spatial ability (Hyde & Linn, 1988). Sex differences in verbal fluency are apparent early in development (Halpern & Lamay, 2000), and are moderate in size (Hines, 1990). We selected phonological verbal fluency for this purpose as it is a widely used cognitive measure in psychological research. We also included a synonym generation task, which requires participants to generate words that are similar in meaning (associational fluency). Sex difference researchers have also found large sex differences in reading comprehension and writing (Lynn, 1992), and so we also included a measure of reading and grammatical skills known to produce moderately large sex differences (Stanley, Benbow, Brody, Dauber, & Lupkowski, 1992).

![Fig. 1. Nash’s (1979) sex-role mediation theory of cognitive abilities.](image)
Finally, we administered sex-role instruments to allow a test of the sex-role mediation hypothesis. The primary focus was on personality traits as measured by the Bem Sex-Role Inventory. This instrument provides a masculine and feminine scale, which on the basis of a median split classifies participants into one of four categories: masculine (M), feminine (F), androgynous (high M, high F) and undifferentiated (low M, low F). We also included additional sex-role measures to determine which aspects of sex-role identity (personality traits, behavioral actions, or social identity) best predicted performance on the spatial and verbal ability tasks.

4. Hypotheses

1. Consistent with past research, we hypothesized that males would perform higher than females on all spatial ability measures.
2. Participants with higher masculinity scores (masculine and androgynous groups) will perform better on spatial ability tasks than participants with lower masculinity scores (feminine and undifferentiated), consistent with the sex-role mediation hypothesis (Nash, 1979).
3. Regression analysis with the three sex-role measures will determine which aspect of sex-role identity (personality, behavior, or social identity) is the best predictor of spatial ability.
4. We hypothesized that sex differences in spatial ability are mediated by masculine sex-role identity.
5. Consistent with past research, we hypothesized that females would perform higher than males on all verbal ability measures.
6. Participants scoring high on femininity (feminine and androgynous groups) will perform better on verbal tasks than participants scoring low in femininity (masculine and undifferentiated), consistent with the sex-role mediation hypothesis.
7. Regression analysis with the three sex-role measures will determine which aspect of sex-role identity (personality, behavior, or social identity) is the best predictor of verbal ability.
8. We hypothesize that sex differences in verbal ability are mediated by feminine sex-role identity.

5. Method

5.1 Participants

Three hundred and nine participants (105 males, 204 females) were recruited from a university subject-pool of psychology students, currently completing a research methods course. The majority of these students were enrolled in an undergraduate psychological science degree. The remainder were enrolled in other health or science programs in which the research methods course was required or recommended as an elective. The mean age was 25.46 years (SD = 8.03), and there was no significant difference in age between male and female participants, \( t(304) = 1.21, p = .228 \). As the distribution of psychological sex-roles is not even in college samples, recruiting a larger number of participants was necessary to ensure a reasonable cell size for analysis in each of the four sex-role categories. All participants provided informed consent to a protocol approved by the Institutional Human Research Ethics Committee.

5.2 Sex-role measures

The 30 item short form of the Bem Sex-Role Inventory (BSRI; Bem, 1974, 1981a) was used to measure sex-role identity. The BSRI is a general personality inventory that incorporates 10 masculine, 10 feminine, and 10 neutral personality traits and items to detect social desirability bias. Each item is rated on a 7-point Likert scale (from “1 — never or almost never true of me” to a midpoint of “4 — occasionally true” and to “7 — always or almost always true of me”). Separate masculine and feminine scores are produced by averaging responses across scale items, resulting in two continuous variables for use in regression analysis. Participants can also be categorized on the basis of a median split into one of four sex-role categories, masculine, feminine, androgynous (high in masculinity and femininity) and undifferentiated (low in both masculine and feminine traits).

With the passage of time since the original publication of the Bem Sex-Role Inventory, it is possible that prevailing gender norms and values may have shifted in the intervening period. If so, the gendered nature of its test items may not reliably discriminate between the constructs of masculinity and femininity for modern samples. Choi, Fuqua, and Newman (2007) investigated the factor structure of the BSRI in a college sample, finding support for distinct masculine and feminine factors. Similar findings emerged in a confirmatory factor analysis across both college and community samples, with the conclusion that the instrument remains valid with modern samples (Choi, Fuqua, & Newman, 2009). For a further discussion on the history and properties of this instrument, see Wood and Eagly’s (2015) review.

A second measure of sex-role identity, the Personal Attributes Questionnaire (PAQ; Spence, Helmreich, & Holahan, 1979) was also used. This is a 24-item self-report measure that includes a mixture of eight stereotypically masculine and eight feminine personality traits. Participants rate themselves on a bipolar 5-point scale (e.g. “very passive” versus “very active”). Although the BSRI and PAQ were highly correlated in our sample \( r = .72 \) for masculinity, \( r = .83 \) for femininity, they represent somewhat different conceptualizations of gender identity and this distinction was used in regression analyses to determine which measure was a stronger predictor of cognitive ability.

We also administered the identity subscale of the Collective Self-Esteem Scale (CSES; Luhtanen & Crocker, 1992) which is a brief four-item measure assessing gender-based social identity. Some researchers make a distinction between gender identity based on sex-typed personality traits, and self-categorization by the individual (Wood & Eagly, 2015). Including a social identity measure makes it possible to examine the relative strength of gender identity associations in predicting intellectual performance. Items include “Overall, I feel that the gender group of which I am a member is not worthwhile” and “In general, I’m glad to be a member of the gender group I belong to”, with two items being negatively coded. Items are rated on a 7-point Likert-type scale ranging from “strongly disagree” to “strongly agree”. Higher scores indicate greater identification with one’s biological gender.

5.3 Spatial cognitive measures

5.3.1. Vandenberg Mental Rotation Task (VMRT)

A computer administered version of the mental rotation task was employed. The stimuli were Vandenberg and Kuse’s (1978) original three-dimensional (3D) stimuli which had been redrawn by Peters et al. (1995). On each trial, participants were presented with a target 3D image and asked to select the two images (from the 4 options) that were rotated images of the target. Response times were measured from onset of the target image until both selections were made. The standard Vandenberg scoring system awards a full mark for locating both of the rotated targets and no mark if one or none was identified (Peters et al., 1995). This scoring method discourages guessing and increases task difficulty. Participants completed a series of practice items with feedback. Participants were allocated 3 min to complete a block of 12 items, followed by a brief rest period and then a second block of 12 items. The time remaining was displayed for each block, and participants were instructed that accuracy was important, as an item would only be scored correctly if both targets were located. The maximum score for the test is 24 (Cronbach’s \( \alpha = .94 \) for the current sample).

5.3.2. 2D Mental Rotation Task (2DMRT)

Previous research has used bar histograms as stimuli to test for sex differences in mental rotation of two-dimensional (2D) stimuli. We employed the same stimuli used in Neumann, Fitzgerald, Furedy, and...
Boyle (2007) as a computer administered task, recording reaction time (in ms) and accuracy. Participants were presented with two bar histograms, and given up to 5 s to correctly identify whether they matched by pressing a key (an error was recorded in the event of non-response). Participants were given practice session containing 15 items, followed by a randomized sequence of 40 actual test items (half of which were matching, half of which did not match). The dependent variable was the accuracy rate expressed as a percentage (current sample Cronbach’s $\alpha = .81$).

5.3.3. Piaget Water Level Test (PWLT)

Spatial perception was assessed using a computer administered version of the Piagetian Water Level Task (reviewed in Vasta & Liben, 1996). Participants were presented with a 2D depiction of a container in the centre of the screen, as well as a flat un-tilted table at the bottom of the screen to represent the horizontal plane. They were instructed that the container would then be tilted (as shown on screen), and that using the computer mouse they should draw the waterline from a starting point on the right side of the container. The stimuli were depictions of equal-sized vessels containing varying volumes of liquid such that the vessel was 20%, 50% or 80% full. The vessels were tilted at angles of 0, 20, 30, 40 or 50° from the horizontal. Participants were administered two trials of each angle (0°, 20°, 30°, 40°, 50°) in a random sequence for a total of 10 trials, with varying heights of liquid (20%, 50% or 80%). The dependent variable for this measure was the average angular error from the horizontal plane across all trials (current sample Cronbach’s $\alpha = .83$). Additionally, participants rated their level of confidence on a 7-point scale for each trial ranging from “1 — not very confident” to “7 — very confident” (current sample’s Cronbach’s $\alpha = .95$).

5.3.4. Group Embedded Figures Test (GEFT)

Spatial visualization was measured using the paper and pencil version of the Group Embedded Figures Test (Witkin, Dyk, Faterson, Goodenough, & Karp, 1962). This task is thought to require respondents to dis-embed a target object from a geometric background, and to isolate distracting stimuli (Linn & Petersen, 1985; Witkin, 2003). Participants are shown a series of complex geometric shapes and asked to locate the target by tracing its outline. After completing a practice item, participants are given four minutes to complete a block of nine items. This was followed by a 1 min rest period and a second block of nine items. Scoring was one point for a correct item, and zero for omitted or incorrect items, for a maximum score of 18 (Cronbach’s $\alpha = .86$ in the current study).

5.4. Verbal cognitive measures

5.4.1. Phonological verbal fluency

Participants are given a letter of the alphabet and asked to generate and write down as many words as possible in 60 s. They were instructed that only unique words were permitted (e.g. if run was given, then runs or running should not be given as answers), and that names of people, brand names, and places would be marked as incorrect. After a practice trial, participants were given four letters in order of F, A, S, and C, one at a time. Internal consistency was high for the sample (Cronbach’s $\alpha = .89$). The average number of words reported over the four trials was the dependent variable.

5.4.2. Synonym generation task

Following the methodology of Hines (1990) and Halpern and Wright (1996), participants were given stimulus words and asked to write as many synonyms as possible within the time limit of 60 s per trial. A practice trial with sample synonyms was presented first to ensure that participants understood the task requirements. The six stimulus items (strong, dark, wild, sharp, turn and clear) were drawn from Hines (1990). Four online dictionaries (Oxford, Collins, Cambridge, and Thesaurus.com) that offered comprehensive definitions of word meanings and lists of synonyms were used to determine whether the reported words were correct as synonyms of the stimulus words. One mark was awarded for each correct synonym. Internal consistency in the current study was high, Cronbach’s $\alpha = .90$.

5.4.3. Differential aptitude test — language usage (DAT-L)

This instrument measures an individual’s ability to detect errors in grammar, punctuation, and capitalization in written text. Participants are presented with individual sentences of text and asked to identify where in the sentence the error is located. To discourage guessing, some items have no error present which the subject must also identify correctly. The test has a multiple-choice format with response options. It contains 30 items and a time limit of 10 min is imposed. As there are regional differences between Australian and American English, the Australian version of the DAT-L was used (Cronbach’s $\alpha = .79$ for the current sample).

5.5. Procedure

Participants were advised that they were participating in a study on cognitive problem solving and personality traits and then undertook either the block of spatial (VMRT, 2DMRT, Piaget WLT, GEFT) or block of verbal tasks (phonological verbal fluency, synonym generation task, DAT-L) with the presentation order of the spatial and verbal task blocks counterbalanced. A rest period of 4 min was given between task blocks to prevent fatigue. In order to minimize gender priming effects, the sex-role personality inventories and demographic information questionnaire were administered after the cognitive testing had been completed. Participants were debriefed and thanked for their participation.

Statistical analysis was conducted using a series of factorial ANOVAs, and in order to avoid Type I error inflation resulting from multiple comparisons, a planned linear contrast was made based a priori on experimental hypotheses. Linear contrasts offer the advantage of increased statistical power by pooling two or more cells. When testing the effect of masculinity on spatial reasoning, a linear contrast compared high masculinity groups (masculine and androgynous) with low masculinity groups (feminine and undifferentiated). Similarly when testing the effect of femininity on verbal reasoning, a linear contrast compared those scoring high on femininity (feminine and androgynous) with those scoring low on femininity (masculine and undifferentiated). Regression analysis also investigated linear associations between sex-role measures and outcomes, overcoming the limitation of small ANOVA cell sizes. Mediation analysis was performed using the PROCESS macro for SPSS (Preacher & Hayes, 2004).

6. Results

6.1. Sex-role classification

Participants were classified into one of four sex-role categories, based on a median-split of BSRI masculinity and femininity scores (Masculinity = 4.50, Femininity = 5.50). Table 1 presents the distribution of males and females classified according to the four sex-roles, while Table 2 presents the descriptive statistics for gender-related

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Distribution of sex-roles in the sample.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Masculine</td>
</tr>
<tr>
<td>Males</td>
<td>38 (36.2%)</td>
</tr>
<tr>
<td>Females</td>
<td>44 (21.6%)</td>
</tr>
</tbody>
</table>
measures. A Chi-square test showed that sex-role classification was dependent on sex of the participants, $\chi^2(3) = 14.09, p = .003$.

$t$-Tests for independent samples showed that masculinity scores were significantly higher for male than female participants, $t(307) = 2.44, p = .015, d = .25$. As would be expected, females also scored higher on femininity scores than males, $t(307) = -3.94, p < .001, d = -.40$. A similar pattern of sex differences was found with the PAQ measures for masculinity, $t(307) = 3.14, p = .002$, and for femininity, $t(307) = -4.74, p < .001$, which were highly correlated with BSRI masculinity ($r = .72$) and femininity ($r = .83$) scores.

Although there were no significant sex differences in CSES, $t(307) = 1.26, p = .210$, there was a statistically significant difference across sex-role categories, $F(3, 305) = 5.76, p = .001, \eta^2_p = .05$, with masculine and androgynous participants scoring higher than feminine and undifferentiated participants, $t(305) = 4.13, p < .001, d = .48$.

### 6.2. Descriptive statistics for spatial and verbal measures

Table 3 presents the descriptive statistics partitioned across sex-role categories for all spatial measures, while Table 4 presents these for verbal measures. Bivariate correlations between all measures are reported in Table 5.

### 6.3. Visuospatial measures

#### 6.3.1. Vandenberg Mental Rotation Task (VMRT)

Performance on the VMRT was analyzed with a 2 (Sex) $\times$ 4 (Sex-Role) factorial ANOVA (see Fig. 2). Significant main effects of sex, $F(1, 298) = 17.26, p < .001, \eta^2_p = .06$, and of sex-role, $F(3, 298) = 5.31, p = .001, \eta^2_p = .05$, were found. The interaction between sex and sex-role was not significant, $F(3, 298) = 1.09, p = .356$. As can be seen in Fig. 2, males scored higher in overall accuracy than females, $t(304) = 4.16, p < .001, d = .48$. The planned linear contrast showed that participants classified as masculine and androgynous scored higher than those classified as feminine and undifferentiated, $t(304) = 3.51, p < .001, d = .40$.

To explore whether the group differences were the result of a speed-accuracy trade off, we also examined reaction times with a 2 (Sex) $\times$ 4 (Sex-Role) factorial ANOVA. There was no significant main effect of sex, $F(1, 298) = 3.17, p = .076, \eta^2_p = .01$. Males and females did not differ in the amount of time spent on items. However, there was a statistically significant main effect of sex-role, $F(3, 298) = 4.65, p = .003, \eta^2_p = .05$ (see Fig. 3). Contrary to that expected by a speed-accuracy trade off, masculine and androgynous participants took more time on the items than the masculine and androgynous participants, $t(304) = 3.35, p = .001, d = .39$. The interaction was not significant.

#### 6.3.2. Mental Rotation Task (2D)

We investigated the percentage accuracy and reaction time (RT) across trials for the Mental Rotation Task (2D) with separate 2 (Sex) $\times$ 4 (Sex-Role) factorial ANOVAs. On inspection of the histogram for percentage accuracy, there appeared to be evidence of a ceiling effect with a large percentage of participants making few or no errors in judgment (median = 90%). As the distribution was extremely negatively skewed, a logarithmic reflection was applied. Contrary to expectations, there was no significant main effect of sex, $F(1, 298) = 1.11, p = .292$, or sex-role, $F(3, 298) = .28, p = .839$, nor was there a significant interaction. Likewise, there were no group differences for reaction times, and no further analysis was undertaken for this task.

#### 6.3.3. Piaget Water Level Task (PWLT)

Angular error and confidence on the PWLT were analyzed with 2 (Sex) $\times$ 4 (Sex-Role) factorial ANOVAs. The distribution of angular errors was positively skewed with some heterogeneity of variance, and accordingly a square root transformation was applied. However as the outcome of all ANOVA tests did not differ, the untransformed data are reported. As predicted there was a significant main effect of sex, $F(1, 298) = 6.62, p = .032, \eta^2_p = .02$, with less angular error for males. There was also the predicted main effect of sex-role, $F(3, 298) = 9.89, p < .001, \eta^2_p = .05$ (see Fig. 3). The interaction between these factors was not significant, $F(3, 298) = 1.16, p = .325$. The planned contrast showed that masculine and androgynous participants showed less angular error than feminine and undifferentiated participants, $t(305) = -.51, p < .001, d = -.58$.

Additionally, males reported greater confidence in estimating the angle than females, $F(1, 298) = 11.94, p = .001, \eta^2_p = .04$, and there was also a significant main effect of sex-role category, $F(3, 298) = 5.13, p = .002, \eta^2_p = .05$. The interaction was non-significant, $F(3, 298) = .39, p = .753$. The planned linear contrast showed that masculine and androgynous participants had higher self-confidence ratings on the Piaget task than feminine and undifferentiated participants, $t(305) = 3.59, p < .001, d = .41$.

#### 6.3.4. Group Embedded Figures Test (GEFT) performance

Performance on the GEFT was analyzed with a 2 (Sex) $\times$ 4 (Sex-Role) factorial ANOVA. Surprisingly, there was no significant main effect for sex, $F(1, 300) = .05, p = .828, \eta^2_p = .00$. As predicted, there was significant main effect of sex-role identity, $F(3, 300) = 4.75, p = .003, \eta^2_p = .05$. Results of the planned contrast showed masculine and androgynous participants scored higher overall for the GEFT than feminine and undifferentiated participants, $t(306) = 3.43, p = .001, d = .39$. The interaction between sex and sex-role was non-significant, $F(3, 300) = .85, p = .467, \eta^2_p = .01$ (see Fig. 4).

#### 6.3.5. Sex-role mediation of spatial ability

In order to perform a more detailed regression analysis of the sex-role mediation hypothesis and minimize the need for multiple comparisons, we first converted scores on each of the spatial tasks into standardized z-scores. Next, because all spatial performance measures were significantly correlated (see Table 5), we calculated the mean standardized score for each participant. The resulting composite score was used as the criterion variable for spatial ability.

We then performed a hierarchical multiple regression on spatial ability (see Table 6). Sex was entered as the sole predictor in Step 1, followed by the two sex-role measures (BSRI and PAQ) for masculinity and femininity, as well as the CSES in Step 2. Although only masculinity was hypothesized to make a contribution to spatial performance,
femininity was included to rule out the possibility of a significant negative association. Assumptions of normality of residuals, linearity of associations, absence of multi-collinearity and homoscedasticity were met. At Step 1, sex made a significant contribution to spatial ability, $F_{\text{reg}}(1, 306) = 14.82, p < .001$, accounting for 4.6% of the variance in spatial ability. Step 2 introduced the sex-role measures, $F_{\text{reg}}(6, 301) = 14.53, p < .001, R^2_{\text{reg}} = .19$, explaining a total of 23.2% of the variance in spatial ability. While both BSRI and PAQ masculinity had significant bivariate correlations, only BSRI made a significant unique contribution ($\beta = .33, p < .001$), with a large portion of the variance in spatial ability being shared by the BSRI and PAQ masculinity measures. Furthermore, there was no significant association between femininity scores and spatial performance. Additionally, there was a small contribution of collective self-esteem to spatial ability, but this failed to achieve statistical significance ($\beta = .10, p = .068$).

Because we had established that masculinity made a significant contribution even after controlling for the effect of biological sex, we next sought to test whether sex-role identity was acting as a statistical mediator. Baron and Kenny (1986) offered a formal set of criteria for testing statistical mediation. Firstly there should be a significant association between the predictor variable (biological sex) and the outcome (spatial ability) ($\beta = - .22, p < .001$). Secondly the relationship between the predictor and the mediator (masculine sex-roles) was significant ($\beta = - .14, p = .015$, Path A). Thirdly the association between the mediator and the outcome should still be significant after controlling for the effect of sex ($\beta = - .41, p < .001$, Path B), as shown in Fig. 5.

We further tested this model using the Sobel test which was also significant, $Z = -2.25, p = .025$ and calculation of the bootstrapped estimate showed that it differed significantly from zero in that the confidence intervals did not span zero (95% CI = −.16 to −.02).

Having established these necessary preconditions for mediation, we tested whether the relationship was partially or fully mediated. In a full mediation model, the association between predictor variable and the outcome will become zero and non-significant after controlling for the effect of the mediator (Path C). If the predictor variable still makes a direct contribution to the outcome even after controlling for the mediator, it can be said to be only partially mediated. Though the beta weight was significantly diminished after controlling for the mediator, there was still a significant association between biological sex and spatial ability ($\beta = - .16, p = .002$). In support of the sex-role mediation hypothesis, the relationship between sex and spatial ability was partially mediated by sex-roles, but sex also made a direct contribution to spatial ability.

### 6.4. Verbal and language measures

#### 6.4.1. Phonological verbal fluency task

The mean number of words reported across trials was analyzed with a 2 (Sex) × 4 (Sex-Role) factorial ANOVA (see Fig. 6). We did not find the expected main effect of sex in our sample, $F(1, 298) = .07, p = .785, \eta^2_g = .00$. However, the predicted main effect of sex-role was found, $F(3, 298) = 5.57, p = .001, \eta^2_g = .05$. Planned contrasts showed that participants classified as having masculine and undifferentiated sex roles wrote fewer words than those classified as having feminine and androgynous sex-roles, $t(304) = -4.03, p < .001$, $d = - .46$. The interaction between sex and sex-role fell short of statistical significance, $F(3, 298) = 2.46, p = .063, \eta^2_g = .02$.

#### 6.4.2. Synonym generation task

The total number of synonyms generated across trials was analyzed with a 2 (Sex) × 4 (Sex-Role) factorial ANOVA (see Fig. 7). There was slight positive skewness, so a square root transformation was applied before analysis. This did not change the outcome, so the results of the ANOVA on the untransformed data are reported. As with the phonological fluency task, the predicted main effect of sex was not found, $F(1, 298) = .29, p = .592, \eta^2_g = .00$. However, there was the expected main effect of sex-role category, $F(3, 298) = 5.65, p = .001, \eta^2_g = .05$.

### Table 3

<table>
<thead>
<tr>
<th>Spatial measure</th>
<th>Masculine</th>
<th>Feminine</th>
<th>Androgynous</th>
<th>Undifferentiated</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMRT RT (ms)</td>
<td>16,163</td>
<td>3454</td>
<td>21,924</td>
<td>8144</td>
<td>17,807</td>
<td>4752</td>
<td>17,443</td>
<td>5165</td>
<td>18,404</td>
<td>5057</td>
<td>22,391</td>
<td>8616</td>
<td>18,312</td>
<td>5248</td>
</tr>
<tr>
<td>PWLT confidence</td>
<td>5.77</td>
<td>3.99</td>
<td>10.44</td>
<td>5.13</td>
<td>11.95</td>
<td>4.55</td>
<td>10.44</td>
<td>5.13</td>
<td>11.95</td>
<td>4.55</td>
<td>10.44</td>
<td>5.13</td>
<td>11.95</td>
<td>4.55</td>
</tr>
</tbody>
</table>

Note: VMRT = Vandenberg Mental Rotation Task; 2DMRT = 2D Mental Rotation Task; PWLT = Piaget Water Level Task; GEFT = Group Embedded Figures Test.

### References


D. Reilly et al. / Intelligence 54 (2016) 147–158.
with no interaction between sex and sex-role, \( F(3, 298) = .23, p = .875 \), \( \eta^2_p = .00 \). Feminine and androgynous participants generated more synonyms than masculine and undifferentiated participants, \( t(304) = 3.89, p < .001, d = .45 \).

### 6.4.3. Differential aptitude test — language usage

Overall performance on the DAT was analyzed with a 2 (Sex) × 4 (Sex-Role) factorial ANOVA (see Fig. 8). As predicted, there were significant main effects of sex, \( F(1, 298) = 3.97, p = .047, \eta^2_p = .01 \), and of sex-role, \( F(3, 298) = 4.91, p = .002, \eta^2_p = .05 \). Planned contrasts showed females had higher scores than males, \( t(304) = -1.99, p = .047, d = -.23 \), while feminine and androgynous participants scored higher than masculine and undifferentiated participants \( t(304) = -3.80, p < .001, d = -.44 \). The interaction between sex and sex-role fell short of statistical significance, \( F(3, 298) = 2.12, p = .097, \eta^2_p = .02 \).

### 6.4.4. Sex-role mediation of verbal ability

In order to test statistical mediation for verbal ability, we converted performance on verbal measures into standardized z-scores. As all verbal measures were significantly correlated (see Table 4), we calculated the mean standardized score for each participant across the three verbal measures and used this as the criterion variable. Next we performed a hierarchical regression analysis to determine which measures of sex-role identity best predicted performance on language tasks (see Table 7). Assumptions of normality of residuals, linearity of associations, absence of multicollinearity and homoscedasticity were met. Sex was entered as the sole predictor in Step 1, \( F_{\text{adj}}(1, 306) = 7.82, p = .005 \), explaining approximately 2.5% of the variance in language ability. At Step 2, we added the BEM and PAQ measures of sex-roles, as well as CSES. Although only femininity was hypothesized to make a contribution to language performance, masculinity was included in the regression model to rule out the possibility of a significant negative association. The revised model explained 23% of the variance in language ability, \( F_{\text{adj}}(5, 301) = 16.05, p < .001, R^2_{\text{adj}} = .21 \). The BSRI femininity scale \( (\beta = .25, p = .007) \) and the PAQ femininity scale \( (\beta = .20, p = .029) \) each made a significant unique contribution, although there was a considerable overlap between these instruments. BSRI femininity was the stronger predictor, so it was used in the mediation analysis. Importantly, masculinity did not contribute to language ability. Additionally, the CSES made a significant contribution \( (\beta = -.16, p = .004) \), with lower scores associated with better performance on the verbal tasks.

Having established that femininity made a significant contribution to verbal ability, we tested for statistical mediation using the BSRI femininity measure (see Fig. 9). First, there was a significant association between sex and verbal ability \( (\beta = .16, p = .005) \). Second, there was a significant association between sex and the mediator variable of femininity \( (\beta = .22, p < .001, \text{Path A}) \). Next, there was a significant association between the mediator (femininity) and verbal ability after controlling for sex \( (\beta = .40, p < .001, \text{Path B}) \). Results of the Sobel test show a statistically significant mediation model, \( Z = 3.48, p < .001 \), and calculation of the bootstrapped estimate of the indirect effect showed that it differed significantly from zero (95% CI = .07 to .27). After controlling for the mediator, the direct effect of sex on verbal ability was no longer significant \( (\beta = .07, p = .195) \), yielding evidence for a full mediation model (see Fig. 9).

### Table 4

<table>
<thead>
<tr>
<th>Measure</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonological Verbal Fluency</td>
<td>10.92</td>
<td>3.36</td>
<td>14.42</td>
<td>1.84</td>
<td>11.91</td>
<td>3.02</td>
<td>11.66</td>
<td>3.65</td>
</tr>
<tr>
<td>Synonym Generation</td>
<td>3.26</td>
<td>1.33</td>
<td>4.94</td>
<td>2.08</td>
<td>3.76</td>
<td>1.90</td>
<td>3.34</td>
<td>1.61</td>
</tr>
<tr>
<td>Language Usage</td>
<td>18.71</td>
<td>5.60</td>
<td>23.67</td>
<td>4.44</td>
<td>22.65</td>
<td>5.80</td>
<td>19.51</td>
<td>5.50</td>
</tr>
</tbody>
</table>

### Table 5

Bivariate correlations between sex, sex-roles, and cognitive measures of spatial and verbal abilities.

| Measure                          | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 11   | 12   | 13   | 14   | 15   | 16   |
|----------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 1. Biological sex                | -.14 | .22  | .26  | -.07 | -.29 | .17  | .17  | -.21 | .01  | .06  | .11  | .21  | -.22 | .16  |     |     |
| 2. BSRI masculinity              | -.01 | .72  | .25  | -.29 | -.29 | -39  | .40  | .30  | -.04 | -.16 | -.08 | .43  | -.13 |     |     |     |
| 3. BSRI femininity               | -.07 | .83  | .03  | -.13 | .12  | .03  | .04  | .33  | .36  | .29  | .05  | .42  |     |     |     |     |
| 4. PAQ masculinity               | -.14 | .33  | .26  | -.22 | -.38 | .45  | .24  | -.04 | -.20 | -.10 | .39  | -.15 |     |     |     |     |
| 5. PAQ femininity                | -.02 | .13  | .12  | .04  | .00  | .01  | .32  | .37  | .32  | .07  | .43  |     |     |     |     |     |
| 6. Collective Self-Esteem        | -.17 | .22  | -.18 | .25  | .14  | .16  | -.27 | .03  | -.22 | -.17 |     |     |     |     |     |     |
| Scale                           | -    | -.40 | -.37 | .41  | .30  | .01  | .06  | .13  | .74  | .06  |     |     |     |     |     |     |
| 7. VMRT score                   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| VMRT time                       | -.14 | -.19 | -.12 | .21  | .09  | .15  | -.29  | .19  |     |     |     |     |     |     |     |     |
| 9. Piaget WLT Angular Error     | -.02 | -.35 | -.01 | .08  | -.07 | .77  | .01  |     |     |     |     |     |     |     |     |     |
| Piaget WLT Confidence Rating    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 11. Group Embedded Figures Test | -.08 | .05  | .18  | .73  | .13  |     |     |     |     |     |     |     |     |     |     |     |
| 12. Verbal fluency              | -.61 | .43  | .05  | .84  | .03  |     |     |     |     |     |     |     |     |     |     |     |
| 13. Synonym generation          | -.37 | .01  | .82  |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 14. Differential aptitude test  | -.17 | .76  | .08  |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Language usage                  |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Spatial composite score         |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 16. Verbal composite score      |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
7. Discussion

The present study provided a more comprehensive test of Nash’s (1979) sex-role mediation hypothesis than has been previously conducted, using a range of measures tapping aspects of spatial reasoning and verbal ability. The expected sex differences emerged for some but not all tasks. However, sex-role differences were found for most spatial and all verbal tasks. Masculine and androgynous participants performed better on spatial tasks than feminine and undifferentiated participants. Feminine and androgynous participants performed better on language tasks. Despite social and political changes in the roles of men and women in the intervening period, support for the sex-role mediation hypothesis was found in a modern sample of college students.

7.1. Spatial ability

Predicted sex differences in performance were found for 3D mental rotation and spatial perception tasks, but not for spatial visualization as assessed by the GEFT, nor for mental rotation as measured by the 2D task. In a meta-analysis of sex differences in spatial ability, Voyer et al. (1995) found that spatial visualization effect sizes were considerably smaller than for other spatial tasks and not consistently found across all studies. Inclusion of the GEFT in future studies with larger samples may be warranted to determine whether sex differences in spatial visualization are still reliably found in modern samples.

The absence of a sex difference for the novel 2D rotation task was surprising, as previous researchers had documented substantial sex differences in mental rotation tasks with similar stimuli (Collins & Kimura, 1997; Prinzel & Freeman, 1995). We note, however, that a previous study by Jansen-Osmann and Heil (2007) found that by manipulating the task difficulty level, sex differences on rotation tasks can be substantially reduced or even eliminated. Therefore, a failure to find any group differences may well be due to the ease of our version and a failure to differentiate between high- and low-ability as indicated by the ceiling effect (median accuracy > 90%). Sex differences in performance were observed in the 3D Mental Rotation Task, and males completed items on this task more quickly than females. This may reflect greater confidence for males on spatial tasks, which was found for explicit confidence ratings on the Piaget Water Level Task.

While previous research had established support for the sex-role mediation hypothesis with some types of spatial ability (e.g., mental rotation), it was unclear whether this effect would generalize to other forms of spatial reasoning. Our study found robust sex-role differences across 3D mental rotation, spatial perception, and spatial visualization tasks, consistent with Nash’s hypothesis. Regression analysis of a composite spatial ability score confirmed a significant association with masculine sex-roles, and importantly, that a corresponding negative association with feminine sex-roles was not present. When all three gender-related measures were entered into the regression model, BSRI masculinity emerged as the strongest predictor of spatial performance, though there was considerable overlap with other measures. Mediation analysis found that masculine sex-role identity was a significant mediator of the relationship between sex and spatial ability.

<table>
<thead>
<tr>
<th>Step</th>
<th>β</th>
<th>t</th>
<th>p-Value</th>
<th>sr²</th>
<th>R</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-22</td>
<td>-3.85</td>
<td>&lt;.001***</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-15</td>
<td>-2.93</td>
<td>.004**</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>.33</td>
<td>4.23</td>
<td>&lt;.001***</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-.11</td>
<td>-1.19</td>
<td>.227*</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>1.38</td>
<td>-170</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>.12</td>
<td>1.33</td>
<td>.186</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>1.83</td>
<td>.068</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 2. Mean performance of males and females on the Vandenberg Mental Rotation Task. Error bars indicate ± 1 S.E.M.

Fig. 3. Mean angular error on the Piaget Water Level Task for males and females. Error bars indicate ± 1 S.E.M. graph.

Fig. 4. Mean performance on the Group Embedded Figures Test (GEFT) for males and females. Error bars indicate ± 1 S.E.M. graph.
results of regression analysis qualified that this relationship supported a partial mediation model, but that sex also makes a direct contribution to spatial performance independent of the mediator.

While the primary focus of this study was on cognitive performance as reflected in accuracy and error rates, we also measured reaction time for rotation and confidence ratings for the Piaget Water Level Task. Some researchers have suggested that speed of processing is an important factor in explaining sex differences in mental rotation (Voyer, 2011). However, males and females did not differ in the average amount of time spent on items. Interestingly, we did notice group differences across sex-role categories, with masculine and androgynous participants completing problems faster. Confidence ratings for the Piaget WLT were also higher for males and those high in masculinity, in line with the smaller angular error for these groups.

7.2. Verbal ability and language skills

In contrast to previous studies, we observed no sex differences in verbal fluency or synonym generation for our sample. This is surprising, given the appreciable effect sizes reported in past research (Halpern & Tan, 2001; Hines, 1990). However, significant differences across sex-role categories were found for these tasks, with those scoring high on femininity (feminine and androgynous groups) generating significantly more words than the masculine and undifferentiated groups.

We did, however, observe a meaningful sex difference in grammar and language usage. This is in line with previous research finding small to medium effect sizes (Stanley et al., 1992). Additionally, sex-role differences were considerably larger than the difference between males and females. It also highlights the utility of examining the effect of sex-role identity on cognitive measures, as a comparison of males and females alone would have been unable to detect any group differences in performance for some verbal measures.

Regression analysis showed that all three sex-role measures accounted for unique variance in the composite measure of verbal ability. However, the BSRI (which measures sex-typed personality traits) was the strongest predictor of verbal performance. We conducted mediation analysis, which showed that the relationship between sex and verbal ability was fully mediated by feminine sex-roles. As the distribution of sex-role categories frequently varies from sample to sample, this may explain fluctuations in the magnitude of sex differences across studies. If our sample of males was somewhat lower in femininity and our females somewhat higher, we may well have found a significant sex difference in verbal fluency and synonym generation.

7.3. Implications and limitations

The sex-role mediation hypothesis proposes an additional developmental factor to explain the emergence of sex differences in spatial and verbal abilities: namely that the degree to which individuals identify with masculine and feminine sex-roles may influence their acquisition of spatial and verbal skills, respectively. This, in turn, may influence broader psychological factors such as intelligence.

Some males and females develop a gender-congruent sex-role identity that leads to a restriction of their interests and behaviors (Bem, 1981b; Martin & Ruble, 2004), while others incorporate an androgynous...
Further research is required to investigate the causal mechanisms by which sex roles mediate intellectual development, and to rule out the possibility that sex role identity acts as a proxy for some as yet unspecified factor. While the study presents a test of sex-role mediation in a sample of adult undergraduate students, such a sample may differ from the general population in demographic characteristics such as socioeconomic status and level of education. There is tentative evidence that socioeconomic status moderates the sex difference in visuospatial reasoning at least (Levine, Vasilyeva, Lourenco, Newcombe, & Huttenlocher, 2005), and replication of sex-role mediation in community samples would represent a stronger test of such hypotheses. It may also capture a broader range of sex role attitudes, as cross-sex-typed subjects such as feminine males occur less frequently in the general population and were small in number in our sample. A limitation of this study is the sample size (particularly of males) and the extent to which the results can be generalized to the population of interest. Nevertheless the sex differences observed in our study (i.e., male advantage in spatial tasks and a female advantage on verbal tasks) are broadly in line with those observed in other studies that include large samples.

8. Summary

The findings of our study provide support for Nash's (1979) sex-role mediation hypothesis for both spatial reasoning and verbal language skills in a modern adult sample, despite the passage of time since it was first proposed. Masculine sex-roles were associated with better spatial ability for all three categories of spatial reasoning (mental rotation, spatial perception, and spatial visualization). Feminine sex-roles were associated with better verbal ability (phonological verbal fluency, synonym generation, and grammar and language usage). Regression analysis showed that sex-typed personality traits were the strongest predictor of spatial reasoning and language skills, and that sex-role identity mediated the relationship between sex and spatial/verbal reasoning. It also highlights the utility of measuring sex-role identity for explaining individual differences in specific cognitive abilities, as sex-role differences were found across measures even when sex differences were absent. Further research is required to examine the causal mechanisms by which sex-roles mediate intellectual functioning.

Acknowledgments

This research was supported in part by a Griffith University postgraduate research scholarship.

References


Table 7
Hierarchical multiple regression of verbal ability (N = 308).

<table>
<thead>
<tr>
<th>Variable</th>
<th>β</th>
<th>t</th>
<th>p-Value</th>
<th>R²</th>
<th>R²*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (0 = male)</td>
<td>-.16</td>
<td>-2.80</td>
<td>.005**</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (0 = male)</td>
<td>-.03</td>
<td>.58</td>
<td>.563</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>BSRI masculinity</td>
<td>-.05</td>
<td>-.64</td>
<td>.526</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>BSRI femininity</td>
<td>.25</td>
<td>2.72</td>
<td>.007**</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>PAQ masculinity</td>
<td>-.01</td>
<td>-.14</td>
<td>.893</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>PAQ femininity</td>
<td>.20</td>
<td>2.20</td>
<td>.029</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>CSES</td>
<td>-.16</td>
<td>-2.91</td>
<td>.004**</td>
<td>.02</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05.
** p < .01.
*** p < .001.

sex-role identity affording greater cognitive and behavioral flexibility (boys and girls can do almost anything). While children receive messages about the suitability and gender-appropriateness of behavior from a variety of sources including peers and culture, parents and educators can have profound influence on the socialization of sex-roles (Witt, 1997). The sex-role mediation hypothesis suggests that there may be tangible benefits for cognitive development in nurturing the interests and talents of both boys and girls who have androgynous flexibility, while gently encouraging those that show a more restricted range of pursuits to diversify their interests. Children who express an interest in educational toys or leisure activities that promote spatial or language development (Baenninger & Newcombe, 1995; McGeown et al., 2011), endorsement of gender stereotypes about the sex-typing of school subjects (Liben, Bigler, & Krogh, 2002), and reduced self-efficacy beliefs and achievement motivation for sex-typed cognitive tasks (Choi, 2004). Stereotype threat might also be a factor (Steele, 1997; Steele, Spencer, & Aronson, 2002), with poorer performance on stereotypically masculine and feminine tasks. While the findings of this study provide support for the hypothesis that sex-roles act as a mediator for the development of sex differences in spatial and verbal abilities, the evidence presented is correlational which by itself cannot prove direct causation. Furthermore, one cannot rule out the possibility that acquisition of sex-role identification may stem in part from children's observations of their own performance in sex-typed cognitive domains such as English and mathematics. Although research shows that children acquire considerable knowledge of sex-roles between 2 and 5 years old, a period that predates formal instruction in subjects like mathematics and reading (Ruble, Martin, & Berenbaum, 2006), it remains possible that competencies for intellectual tasks help further refine one's sex-role identity, or that there are bidirectional links between sex-role identity and intellectual abilities.

Fig. 9. Indirect effect of sex on verbal ability, with feminine sex-roles acting as a mediator.


PloS One, 7(7), e39904. http://dx.doi.org/10.1371/journal.pone.0039904.


the neutral priming condition, a briefing statement was provided that described the wide
variability in performance levels on visual-spatial and verbal tasks, and that the
experiment would be investigating the contribution of personality traits to individual
differences. As before, multiple choice questions followed the briefing statement to
verify that the briefing statement had been read and understood, and in the event of
incorrect answers they were instructed to read the briefing statement again (0.8% of
participants). All participants completed the mental rotation and verbal fluency tasks
with the presentation order counterbalanced to avoid sequencing effects.

After completing the cognitive tasks, participants were presented with a gender
beliefs checklist and the BSRI personality inventory. These items were administered
last to avoid gender-priming in the neutral condition of the experiment, as previous
studies had found that even asking demographic questions about gender can lower the
performance of women on visual-spatial tasks (McGlone & Aronson, 2006; Ortner &
Sieverding, 2008).

Results

Labelling of Spatial Task

Performance on the Group Embedded Figures Test was analyzed using a $2 \times 4$ (Condition: Spatial or Empathy labelling) $\times 4$ (Sex-Role: Masculine, Feminine,
Androgynous, Undifferentiated) factorial ANOVA. Cell sizes for the spatial condition
were Masculine (11), Feminine (15), Androgynous (14), Undifferentiated (30), while
those for empathy condition were: Masculine (23), Feminine (13), Androgynous (24)
and Undifferentiated (20). The assumptions of normality and homogeneity of variance
were met. As shown in Figure 2 and consistent with hypotheses, there was a significant
main effect of condition, $F(1, 142) = 30.88, p < .001, \eta_p^2 = .18$, with a large effect size.
When the task was described as a test of empathy and perspective taking, women scored
significantly higher on the spatial visualization task than when the task was described as being visual-spatial in nature $t(148) = 5.56, p < .001, d = .91$. Furthermore there was a significant main effect of sex-role category, $F(3, 142) = 3.01, p = .032, \eta_p^2 = .06$. To evaluate our hypothesis, we conducted a planned linear contrast to compare those scoring high in masculine sex-role identification (masculine and androgynous groups) with those scoring low in masculine identification (feminine and undifferentiated groups). The planned contrast showed that masculine and androgynous participants scored significantly higher than those classified as feminine and undifferentiated, $t(148) = 3.00, p = .003, d = .48$. The interaction term approached but fell short of statistical significance, $F(3, 142) = 2.57, p = .057, \eta_p^2 = .05$.

**Figure 9.2.** Mean performance on the Group Embedded Figures Test across condition and Bem Sex Role Inventory (BSRI) sex-role categories. Error bars indicate ± 1 S.E.M.
**Stereotype-Threat**

Performance on the mental rotation task was analyzed using a $2 \times (\text{Condition: stereotype-threat or control}) \times 4 (\text{Sex-Role: masculine, feminine, androgynous, undifferentiated})$ factorial ANOVA. Cell sizes for the stereotype-threat condition were Masculine (15), Feminine (18), Androgynous (17), Undifferentiated (24), while those for the control condition were: Masculine (19), Feminine (10), Androgynous (21) and Undifferentiated (26). The assumptions of normality and homogeneity of variance were met for the distributions. As shown in Figure 3, there was a significant main effect of condition, $F(1, 142) = 10.66, p = .001, \eta^2_p = .07$, with women in the stereotype-threat condition scoring significantly lower than women in the control condition $t(148) = .$

There was also a significant main effect of sex-role category, $F(3, 142) = 7.34, p < .001, \eta^2_p = .13$. The planned linear contrast showed that masculine and androgynous women scored higher than feminine and undifferentiated, $t (148) = 3.29, p = .001, d = .54$. The interaction term between condition and sex-role was not significant, $F(3, 142) = .94, p = .423, \eta^2_p = .02$. 
Figure 9.3. Mean performance on the Vandenberg Mental Rotation Task across condition and Bem Sex Role Inventory (BSRI) sex-role categories. Error bars indicate ± 1 S.E.M.

We also examined whether there was an effect of sex-role or condition on the average completion time of all questions answered (displayed in Figure 4). Consistent with previous studies there was a significant main effect of sex-role category, $F(3, 142) = 5.42, p = .001, \eta^2_p = .10$, with the planned contrast showing that masculine and androgynous women completed items faster than feminine and undifferentiated, $t(145) = 3.04, p = .003, d = .50$. There was no main effect of condition on speed, $F(1, 142) = 3.57, p = .061, \eta^2_p = .03$. Interestingly, the interaction between condition and sex-role category was significant, $F(3, 142) = 3.53, p = .016, \eta^2_p = .07$. In order to examine the interaction further, two separate one-way ANOVA’s were conducted. For those participants in the stereotype-threat condition, there was a significant difference across
sex-role categories, $F(3, 71) = 6.67, p < .001, \eta^2 = .22$, with the planned contrast showing that masculine and androgynous participants were faster to complete questions than feminine and undifferentiated participants, $t(71) = -3.93, p < .001, d = -.93$. However there was no significant difference in speed across sex-role categories under the control condition, $F(3,71) = .81, p = .494, \eta^2 = .03$.

![Figure 9.4](image.png)

Figure 9.4. Mean completion time for the Vandenberg Mental Rotation Task items across condition and Bem Sex Role Inventory (BSRI) sex-role categories. Error bars indicate ± 1 S.E.M.

**Stereotype Lift**

Verbal fluency scores were analyzed using a $2 \times$ (Condition: stereotype-lift or control) $\times 4$ (Sex-Role: masculine, feminine, androgynous, undifferentiated) factorial ANOVA. Cell sizes for the stereotype condition were Masculine (15), Feminine (18),

Androgynous (17), Undifferentiated (24), while those for the control condition were: Masculine (19), Feminine (10), Androgynous (21) and Undifferentiated (26). All assumptions of the factorial ANOVA were met. As shown in Figure 5, and contrary to the stereotype-lift hypothesis, women in the stereotype priming condition did not generate more words than those in the control condition, $F(3, 142) = 0.11, p = .739, \eta_p^2 = .00$, nor was the interaction term between condition and sex-role significant, $F(3, 142) = 0.52, p = .667, \eta_p^2 = .01$. However there was a significant main effect of sex-role category, $F(3, 142) = 3.20, p = .025, \eta_p^2 = .06$. The planned linear contrast showed that feminine and androgynous women (high femininity) generated more words than masculine and undifferentiated women (low femininity), $t(148) = -2.91, p = .004, d = -.47$. 
Figure 9.5. Mean number of words generated in verbal fluency task across condition and Bem Sex Role Inventory (BSRI) sex-role categories. Error bars indicate ± 1 S.E.M.

Manipulation Check

As a manipulation check, we tested whether there were any differences in gender beliefs about cognitive abilities between those receiving the stereotype-threat induction and those in the control condition following the briefing. Those in the stereotype-threat condition showed greater endorsement than those in the control condition for visual-spatial ability, $t(147) = 3.99, p < .001, d = .65$, and for verbal language ability, $t(147) = 2.99, p = .003, d = .49$. In addition, to rule out the possibility of a carry-over effect of the experimental manipulation of task-labelling on the GEFT from the first phase of the experiment, we checked whether those assigned to the empathy condition differed in their mental rotation and verbal fluency scores from those assigned to the spatial
condition. Condition was dummy-coded with 0 being spatial condition and 1 being empathy condition. There was no significant difference in mental rotation, $t(148) = -1.74, p = .083$, or verbal fluency, $t(148) = .62, p = .533$, between conditions to confirm that there were no carry-over effects.

**Discussion**

Previous studies on the role of sex-role identification have found differences in performance between high and low masculinity participants on visual-spatial tasks (Reilly & Neumann, 2013). However, the research literature was unclear on whether this reflected enduring differences in sex-role identification (an enduring personality trait), or was a more transient state resulting from situational and intrapersonal factors. Two possible sources of temporary performance decrements were investigated in this experiment, that of task-labelling and stereotype-threat.

**Sex-typing of cognitive tasks and task-labelling**

Firstly, we examined whether the way in which visual-spatial tasks were described to test-takers might influence women’s performance on the GEFT. When the GEFT was portrayed as requiring stereotypically ‘masculine’ traits of visual-spatial reasoning, women scored lower than when the same task was portrayed as requiring stereotypically ‘feminine’ traits of empathy and perspective taking. This is consistent with the earlier study by Brosnan (1998). Brosnan (1998) attributed the effect to the perceived sex-typing of the task and whether it was gender-appropriate. It is possible that women in the empathy condition simply exerted more effort in the face of challenging content than those who believed it to be a masculine sex-typed task. There was also an independent effect of sex-role category in line with the sex-role mediation hypothesis, and which is consistent with two previous studies employing the GEFT in college-aged samples (Brosnan, 1998; Reilly et al., 2016). Regardless of how the spatial
visualization task was described, masculine and androgynous women still scored significantly higher than feminine and undifferentiated women. If the group differences were *solely* a result of the perceived sex-typing of the task, there should be no significant main effect of sex-role category. From these results, it would appear that task labelling and sex-role identity both contributed to task performance, with task labelling producing a temporary effect that added to the sex-role differences which occurred regardless of condition.

Our results differed somewhat from that observed by Massa et al. (2005), who found no overall main effects of task description or sex-role identity in women but did observe a significant interaction between description and sex-role. In our sample, the interaction term fell short of statistical significance, which may reflect sampling variance or differences in study methodology. The procedure employed by Massa et al. differed in that half their participants completed the Bem Sex Role Inventory *before* attempting the spatial visualization task, and half completed the BSRI survey afterwards. This may have led to an unanticipated gender priming effect in some participants, particularly when using the longer form of the BSRI instrument where the gendered nature of the personality traits is more readily apparent. It may also be due to differences between the short and long form of the BSRI instrument, with the short form reportedly possessing better psychometric properties than the longer one (T. Campbell, Gillaspy, & Thompson, 1997; Choi, Fuqua, & Newman, 2009).

**Stereotype Threat/Lift**

We also investigated the effect of negative gender stereotypes about cognitive ability by experimentally inducing stereotype threat. For the mental rotation task, women in the stereotype-threat condition performed more poorly than those in the control condition who were not briefed about patterns of sex difference research for
visual spatial tasks. Stereotype threat is often considered a factor in high-stakes testing (Steele, 1997; Steele & Aronson, 1995), such as performance on standardized tests of aptitude like the SAT for college admission. But our study found a significant stereotype threat effect even for a relatively non-consequential task, in a situation where anonymity was guaranteed. Consistent with past research (Reilly & Neumann, 2013), there was also a significant main effect of sex-role category, with masculine and androgynous groups scoring higher than feminine and undifferentiated groups on mental rotation. There was also a significant interaction between sex-role and condition for reaction times, with masculine and androgynous participants answering questions faster than feminine and undifferentiated participants in the stereotype threat condition but not in the control condition. This may be due to the additional performance pressure caused for the stereotype threat condition.

Several recent studies have experimentally induced the stereotype threat effect for visual-spatial tasks (S. M. Campbell & Collaer, 2009; Heil, Jansen, Quaiser-Pohl, & Neuburger, 2012; McGlone & Aronson, 2006). McGlone and Aronson (Experiment 1a) examined the effect of gender priming on mental rotation performance with the Vandenberg MRT. They found that a manipulation designed to make gender category salient led to poorer performance. They found a medium-sized effect comparable to that observed in our study. A second experiment by Heil, Jansen, Quaiser-Pohl and Neuburger (2012) briefed participants with one of three sets of instructions (either that males were better at visual-spatial tasks, women were better, or that no sex differences exist), followed by a mental rotation test. In their sample, women’s mental rotation performance was poorer when led to believe that men score better on visual-spatial tasks than the neutral condition. Additionally women scored higher than the neutral condition when led to believe that women were better at visual-spatial tasks. Finally,
Campbell and Collaer (2009) also found a stereotype threat effect for a novel visuospatial task involving accuracy of judgment for angular line tasks. The present study replicates these findings on stereotype threat induction, and extends such research to consider the contribution of sex-role identity on performance.

We also sought to test whether the activation of a favourable stereotype (women scoring higher than men on language tasks) might affect performance on a verbal fluency task. When briefed about the strong female advantage of tasks of verbal ability (positive stereotype), women did not generate any more words than those in the control condition as would have been predicted by the stereotype lift hypothesis (Walton & Cohen, 2003). It may be easier to decrease performance with negative gender stereotypes than it is to increase performance with positively affirming stereotypes. Predicted sex-role differences in word production were observed though, consistent with those found previously (Nash, 1979; Reilly et al., 2016).

**Implications and Limitations**

Although a large body of research has demonstrated robust sex differences in cognitive ability for both visual-spatial and verbal domains, researchers disagree on the extent to which they reflect innate biological processes or sociocultural differences between men and women’s roles in society. Hyde (2005) has noted that within-sex variation is often larger than the variability between males and females, and that researchers should investigate factors associated with individual differences in cognitive ability. The sex-role mediation hypothesis proposed by Nash (1979) offers an alternate explanation grounded not in biological sex, but rather in the development of stereotypically masculine and feminine personality traits. It was unclear though whether observed performance differences are due to differences in sex-role identity or to more transient factors associated with the testing situation such as the perceived sex-role
appropriateness of the task or stereotype-threat. The present results for visual-spatial ability suggest that although there are transient effects of the testing situation, there are still significant sex-role differences reflecting an enduring trait. We also found significant sex-role differences in a verbal fluency task, while no difference in fluency was found between women reminded about the strong female advantage in such tasks and the control group.

While many studies have documented the stereotype threat effect in the context of high-stakes testing and educational assessment (such as tests of mathematical achievement), only a few studies (e.g. Massa et al., 2005) have investigated it for other domains such as visual-spatial ability. An important consideration is also the presence of situational cues such as test instructions that might subtly impact on performance. The present results highlight how beliefs about the skills required to complete a challenging cognitive task can serve to boost or impair performance. Even subtle situational cues may inadvertently prime test-takers to think about sex stereotypes, and further research is needed to increase resiliency in such testing situations (Miyake et al., 2010).

The present study recruited a female-only sample because gender priming and stereotype-threat induction have only been demonstrated to exert an effect in women for visual-spatial tasks. There is scant research on similar performance impairments in males on tasks of verbal ability and language, but one recent study has found evidence for stereotype-threat in boys when reading ability is tested (Pansu et al., 2016). Given the moderately large sex differences present on language tasks such as reading, writing, punctuation, and grammar (Halpern, 2011), the question of whether males are similarly susceptible to stereotype threat and whether this interacts with sex role orientation warrants further investigation. Also, while the primary outcome of interest in our study
was performance on visual-spatial and verbal tasks, it is possible that group differences in measured performance might also reflect factors such as motivation and self-efficacy beliefs. Future studies might strengthen the validity of their findings by including such measures.

**Conclusions**

The present study demonstrates the effect of three factors on visual-spatial performance in a young-adult sample of women. Firstly the way in which participants see the task can have a significant effect on their performance. In our experiment women led to believe a task required stereotypically feminine skills of empathy performed better than when instructed it was a test of visual-spatial reasoning. Secondly, we demonstrated that knowledge and priming of gender stereotypes can also lead to diminished performance for visual-spatial tasks, even in the absence of high-stakes testing. In addition to these state effects, there were significant sex-role differences with greater visual-spatial performance by masculine and androgynous women, while feminine and androgynous women showed greater verbal fluency. This pattern of results suggests the sex-role mediation effect observed in previous studies exerts an effect on visual-spatial performance in women, but can be moderated by task-labelling and salience of gender stereotypes.
References


Chapter 10 – Empirical Study 3 - Sex and Sex-Role Differences in Self-Estimated Intelligence (SEI)

“Such is the nature of men, that howsoever they may acknowledge many others to be more witty, or more eloquent, or more learned; yet they hardly believe there be many so wise as themselves.” – Thomas Hobbes, English philosopher.
or “I’m smarter than the average bear” – Yogi Bear

Overview

The link between intelligence and academic achievement is fairly robust (Laidra, Pullmann, & Allik, 2007; Rohde & Thompson, 2007), with correlations between academic achievement and IQ in high school students ranging from .50 and .70 (Jensen, 1998). This effect is even stronger in younger students (Mayes, Calhoun, Bixler, & Zimmerman, 2009). However, a great deal of educational success in post-secondary education depends on non-cognitive factors, such as personality traits, academic motivation, and self-efficacy beliefs. How we see ourselves intellectually – either as smart, academically capable or possessing more mediocre abilities – can have a profound impact on our academic engagement, motivation, and self-efficacy beliefs. These, in turn, guide our intellectual interests and leisure activities, and ultimately the academic and occupational paths we choose (or reject) in later life. But how is our intellectual self-concept formed, and are there sex and sex-role differences in the accuracy of these self-evaluations?

Much of the literature relating to this study has been presented in an earlier chapter (see Sections 2.2.1-2.2.3), but to briefly recap there are several widely observed research findings that have bearing on this issue. Firstly, many people see themselves as being somewhat smarter than the average person, which has been termed the “above-average effect” (Kruger & Dunning, 1999), and rarely do people rate themselves “below average” (McCrae, 1990). The accuracy of self-judgements of intelligence is actually
weak, with correlations between self-reports and psychometrically measured IQ typically falling in the range of $r = .20$ to $.25$ in college samples (Paulhus, Lysy, & Yik, 1998). However a more recent meta-analysis by Freund and Kasten (2012) found slightly higher average effect size of $r = .33$ when including non-academic samples drawn from the community. Secondly, males and females typically differ in their self-estimates, with males providing significantly higher self-estimates of intelligence than females (despite the absence of sex differences in IQ for the general population). This has been termed the “male hubris, female-humility” effect, and has been widely replicated cross-culturally (Furnham, Hosoe, & Tang, 2001; Szymanowicz & Furnham, 2011). While the effect might at least partly reflect a social desirability bias, it is still found in samples drawn from countries which emphasise communal traits of humility and modesty. Thirdly, when rating the intellectual abilities of family members (parents, spouses, and children), male relatives are still given higher scores on average than female relatives by both sexes. Therefore there appears to be a genuine self-enhancing bias in men for self-estimated intelligence and a self-derogatory bias in women (Furnham et al., 2001). Less clear, though, are the mechanisms underlying sex differences in self-estimated intelligence. Research has shown though that these sex differences are found as early as fifth grade, where it is typically labelled as intellectual self-concept (Gold, Brush, & Sprotzer, 1980; Marsh, 1989). The literature differentiates between self-concept and self-esteem (Brinthaupt & Erwin, 1992), with self-concept being the way an individual sees themselves and self-esteem referring to the affective component. However they are strongly intertwined, and a negative intellectual concept over time leads to reduced academic self-esteem.

In addition to overall impressions of global intelligence, sex differences in intelligence are also found for more specific cognitive abilities, such as verbal or
mathematical/scientific reasoning. A large number of studies have presented descriptions of multiple intelligences after the fashion of Gardner (1983, 1999). It is not necessary to accept the underlying premises of Gardner’s theory multiple intelligences, but it does offer an insight into laypeople’s understanding of intelligence and cognitive abilities. That sex differences in estimation of specific cognitive abilities exist may be expected, because of cultural stereotypes associating specific cognitive abilities with gender (Swim, 1994). For example, implicit associations between masculinity and STEM and between femininity and arts/language are widely observed cross-culturally (Nosek, Banaji, & Greenwald, 2002). A meta-analysis by Syzmanowicz and Furnham (2011) found robust sex differences in self-estimated abilities for mathematical/logical ($d = .44$) and spatial intelligence ($d = .43$) for which there are strong prevailing gender stereotypes. However, a significant but much smaller effect was found for self-estimates of verbal intelligence (unweighted $d = .12$), due in part to the presence of four studies where women provided higher estimates than males (ranging from $d = -.38$ to -.15). This might reflect of prevailing gender stereotypes about female proficiency in language having a moderating effect, or prior knowledge in the psychology subject pools sampled of empirical research studies finding a general female advantage on verbal and language abilities (see Section 2.2.1). Few consistent sex differences have been found for other domains of Gardner’s multiple intelligences, though individual studies have reported exceptions (Visser, Ashton, & Vernon, 2008; Yuen & Furnham, 2006).

While the male hubris/female humility effect has been widely documented, a limitation of many previous studies is their reliance on self-reports without assessing intelligence psychometrically. While there is near consensus in the literature that males and females do not differ in intelligence, Becker and Hedges (1988) note that the presence of greater male variability in the population and recruitment of non-randomly
selected samples (such as from a psychology subject pool) may actually yield samples with significant sex differences in intelligence. Without checking for differences in intelligence in convenience samples, it is impossible to determine whether male hubris/female humility effect is genuine or an artefact of sampling bias. Furthermore, participants may be more truthful in providing estimates if they know that they are going to complete an intelligence test, thus minimising social desirability bias.

The present study seeks to address such issues, by investigating sex differences in self-estimated intelligence and concurrently administering the Cattell’s Culture Fair IQ test. Additionally, it will explore whether sex-role identification might explain the apparent differences between males and females, and whether these differences might also be explained by participants’ general self-esteem. The rationale for considering self-esteem was that many people with high self-esteem exaggerate their successes and positive traits to themselves and to others, while people with low self-esteem focus more strongly on failures and negative traits (Baumeister, Campbell, Krueger, & Vohs, 2003). Hansford and Hattie (1982) found a modestly sized correlation between self-esteem and academic performance as measured by GPA, \( r = .34 \), so it does appear that the way in which we see ourselves and intellectual achievement are related (if not necessarily casual). To date, however, only one study has examined the relationship between objectively measured intelligence and self-esteem. Gabriel, Critelli and Ee (1994) found that people with high self-esteem rated themselves as more intelligent than people with low self-esteem (\( r = .35 \)), but the results of the IQ test did not justify such rosy claims because there was no evidence for a relationship between self-esteem and objectively measured intelligence. Additionally, the discrepancy between measured and self-estimated intelligence was also positively correlated with participants self-esteem scores (\( r = .38 \)).
Though not the primary focus of this study, previous literature has also identified associations between sex-role identification and self-esteem, with masculine and androgynous participants reporting high levels of self-esteem and psychosocial wellbeing when compared with feminine and undifferentiated groups (Alpert-Gillis & Connell, 1989; Lau, 1989; Whitley, 1983). Males also report significantly higher levels of general self-esteem than females (Kling, Hyde, Showers, & Buswell, 1999; Major, Barr, Zubek, & Babey, 1999). On this basis it was reasoned that lower self-esteem might partly explain sex and sex-role differences in self-estimated intelligence (SEI).

Hypotheses

1. Consistent with previous research, males will report higher SEI scores for general intelligence than females, even in the absence of objectively evaluated differences in intelligence.

2. Regardless of sex, high masculinity participants (i.e., masculine and androgynous groups) will report higher SEI scores than low masculinity ones (i.e., feminine and undifferentiated).

3. Males and high masculinity groups will report higher general self-esteem and academic self-esteem than females and low masculinity groups.

4. It is hypothesized that sex, masculinity, and general self-esteem will be associated with SEI, even after controlling for psychometrically measured intelligence.

5. Masculinity scores would act as a statistical mediator of the relationship between sex and SEI scores.

6. Sex and sex-role differences will also be found in self-estimates of multiple intelligences, following a similar pattern as observed with general intelligence.
Method

Participants

Two hundred and twenty-eight participants (103 male, 125 female) were recruited from a university subject-pool of students completing a first-year research methods and statistics course. As the distribution of psychological sex-roles is not even in college samples, recruiting a larger number of participants was necessary to ensure a reasonable cell size for analysis in each of the four sex-role categories. This subject pool was chosen because it included psychology and non-psychology students in order to draw from a broader pool of sex-role categories. Sex-role categories are not equally represented in the general population (e.g., feminine-scoring males typically comprise less than 12% of males, while a similar proportion is found for masculine sex-typed females), and recruiting from a broader pool of candidates than just a single discipline (psychology) maximises the likelihood of encountering sufficient numbers for each sex/sex-role category. While the majority of these students were completing an undergraduate psychological science degree (53.7%), a large proportion were enrolled in exercise science or physiotherapy (30.7%), followed by health or biomedical sciences (7.4%) and occupational therapy (4%). Only 3% were studying another type of degree and had selected the subject as an elective. As is typical of university participants, the distribution of ages was strongly positively skewed with a mean age of 22.62 (SD = 6.30, range = 18 to 47 years) and there was no significant difference in age between males and females. Eighty-nine percent of the sample spoke English as their primary language. All participants provided informed consent to a research protocol approved by the Griffith University Human Research Ethics Committee (HREC).
Procedure

Participants were informed that they were participating in a study on the measurement of human intelligence, and the accuracy of self-estimates. They were provided with a booklet containing the self-estimated intelligence (SEI) measures, followed by the Cattell Cultural Fair IQ Test (CCFIT). To prevent intellectual fatigue, rest periods were provided between each subtest of the CCFIT. Following test administration, participants completed a number of personality surveys measuring self-esteem, sex-role identification, and general demographic information. The personality surveys were administered after the self-estimated intelligence survey and CCFIT, in order to minimise gender priming effects on SEI and test performance. Participants were tested in small batches (maximum 3 participants per session) so that compliance with instructions could be monitored and that survey items were read and considered before answering.

Measures

*Self-estimated intelligence (SEI).* Following the methodology of Furnham and Rawles (1995), participants were provided with a simple one page sheet from the booklet which explained in a brief paragraph that the distribution of intelligence in the general population followed a bell curve (see Figure 10.1 for stimuli) that is normally distributed, with the average IQ score being 100 with a standard deviation of 15. This replicated following the methodology employed by Furnham and Rawles (1995) who generously provided copies of stimulus materials, and has been used subsequently in numerous self-estimated intelligence studies. The text of the paragraph was also read aloud by the experimenter to ensure that written instructions were followed. While the properties of the normal distribution were familiar to students in the statistics course, labelled framing anchors were also provided to aid in estimation. Participants were
asked to use this scale to provide an estimate of their intelligence relative to other people, and to write this as a whole number.

![IQ curve]

Modest Impairment      Mild Impairment      Below Average      Average Above Average      Gifted Exceptionally Gifted

Figure 10.1 Stimulus material used for self-estimation of intelligence

On a subsequent page of the booklet, participants read several paragraphs describing research of Gardner’s (1999) theory of multiple intelligences, which defined intelligence more broadly than would be typically assessed by an IQ test. Gardner subsequently revised his model of multiple intelligences to include a total of 9 separate skills (Verbal and linguistic intelligence, Logical-mathematical intelligence, Spatial, Musical, Bodily-kinaesthetic, Interpersonal, Intrapersonal, Naturalistic, and Existential/Spiritual intelligence). Each skill was accompanied by a brief paragraph description that had been pilot tested for readability. An issue identified in pilot testing was that some participants completed the task extremely quickly with minimal variation in scores across domains. So that participants gave considered and deliberated responses, they were instructed to complete the task one definition at a time, and to
record a response *only* after the experimenter had read the paragraph aloud (on the pretence ‘that some participants might come from a non-English background or have reading impairments such as dyslexia, and we want to make sure instructions are clearly understood’). This also ensured that participants had received the appropriate definition for each task, even if they elected not to read the presented material. The definition of existential / spiritual intelligence was phrased for inclusiveness so that it was clear to subjects that this may include but does not require religious practice. Participants responded by providing a numerical IQ score in the same format as for general intelligence.

*Cattell Culture Fair Test of Intelligence* (CCFIT, Cattell, 1973). The CCFIT is a non-verbal measure of fluid intelligence ($g_F$), designed specifically to be as free of culture and educational experiences as possible. This measure was selected for inclusiveness, as it does not require a high standard of English language proficiency and around 10% of the student body come from a background where English is not their primary language. Additionally the student body also includes students who have entered university from alternate pathways (5% of the current sample did not complete high school), or who may have reading difficulties (estimates for Australia range from 5 to over 15% depending on the criteria and sampling methodology (Skues & Cunningham, 2011). It has been designed to minimise cultural or educational biases assumed to be present in the Weschler or Stanford-Binet IQ assessments, by excluding items that require verbal and linguistic proficiency and general knowledge of a specific culture. Additionally, prior research confirmed no sex bias in the CCFIT with equivalent scores for males and females among adult high school graduates (Colom & García-López, 2002).
The specific instrument employed was CCFIT Scale 3, Form A intended for use with adult participants. It contains very clear and simple instructions that are administered in a spoken format by the tester, and requires only a multiple-choice response. The test is also suitable for small group administration, which was desirable to minimise testing time required if participants had been assessed sequentially. The test includes practice items and safeguards to ensure understanding of and compliance with instructions. The instrument also provides appropriate norms tables to allow for conversion between raw scores and their equivalent IQ (centered around a mean of 100 with a standard deviation of 15), for direct comparability to SEI scores provided by subjects. The CCFIT also shows strong convergent validity other tests of general intelligence such as the WAIS with $r = .72$ (Cattell, Krug, & Barton, 1973), and loads highly against more recently revised intelligence scales (Carroll, 1993).

The CCFIT assessment requires inductive reasoning about perceptual patterns, and is comprised of four subtests (series completion, classification, matrices, conditions/typology). Each subtest is completed under strict timing conditions, with items of increasing level of difficulty such that less than 10% of subjects completed all items in the current sample. Although there is no penalty for guessing, two of the subtests require multiple correct responses for the item to be scored correctly. Individual responses were recorded on response sheets that were transcribed and then computer scored for accuracy of scoring. Reliability of the instrument for the current sample was high across the four subtests (Cronbach’s $\alpha = .72$).

General self-esteem. Participants completed the Rosenberg (1965) General Self Esteem Scale, a brief 10 item rating scale that is widely used and demonstrates good psychometric reliability and validity (Sinclair et al., 2010). Participants recorded a response on a 4-point Likert-type scale (ranging from 1- “Strongly Agree”, to 4-
“Strongly Disagree”). Sample items include “On the whole, I am satisfied with myself” and “All in all, I am inclined to feel that I am a failure,”, with several items being reverse coded (Cronbach’s α = .89 for sample).

**Academic self-esteem.** There were two measures. Subjects completed a seven-item Academic Self-Esteem scale (see Appendix A3), adapted for this present study from Johnson et al.’s (1983) Academic Self-Esteem subscale, and Bachman’s (1970) Self-Concept of Ability Scale (SCAS). For comparability, subjects endorsed items on the same 4 point scale used for the Rosenberg GSES. Sample items include “I feel confident in my ability to complete university”, and “I am not doing as well at university as I would like to” with negatively worded items that were reverse coded. Subjects also completed the single item Rosenberg Academic Self-Esteem scale, which asks “How do you rate yourself in academic ability compared with those studying your degree” on a 5-point scale (ranging from 1- “Far below average” to 5- “Far above average”). The final response variable incorporated both measures of academic self-esteem, with high reliability (Cronbach’s α = .87) for the eight-item scale (see Appendix A4).

**Bem Sex-Role Inventory.** The 30 item short form of the Bem Sex Role Inventory (BSRI; 1974, 1981) was used as a measure of sex-role identification. The BSRI is a general personality inventory that includes 10 masculine, 10 feminine as well as 10 neutral and filler items so that the gendered nature of the instrument is not transparent. Traits are rated on a 7-point Likert scale (from “1 – Never or almost never true of me” to a midpoint of “4 – Occasionally true” and ending in “7 – Always or almost always true of me”). Separate masculinity and femininity scores were produced by averaging responses across each scale, resulting in a continuous score suitable for regression analysis. Participants were also categorised on the basis of a median split of their
masculinity ($Mdn = 4.60$) and femininity ($Mdn = 5.30$) scores, to one of four sex-role categories: masculine, feminine, androgynous (high masculinity and high femininity) and undifferentiated (low in both masculine and feminine personality traits). Internal consistency, as assessed by Cronbach’s $\alpha$, was high in the present sample (masculinity scale $\alpha = .81$, femininity scale $\alpha = .85$) and despite the passage of time since its inception the BSRI remains a valid measure of sex-role identification in modern samples (Choi, Fuqua, & Newman, 2007).

**Results**

**Sex Role Classification**

Table 10.1 presents the distribution of sex-role categories in our sample for males and females. As would be expected from past research, an independent samples $t$-test showed that males were significantly higher in BSRI masculinity scores than females, $t(225) = 3.04, p = .003, d = .41$, and that females were significantly higher than males in BSRI femininity, $t(225) = -2.48, p = .014, d = -.33$ than males.

Table 10.1

<table>
<thead>
<tr>
<th>Gender</th>
<th>Masculine</th>
<th>Feminine</th>
<th>Androgynous</th>
<th>Undifferentiated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>29</td>
<td>17</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>(28.2%)</td>
<td>(16.5%)</td>
<td>(33.0%)</td>
<td>(22.3%)</td>
</tr>
<tr>
<td>Females</td>
<td>23</td>
<td>33</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>(18.5%)</td>
<td>(26.6%)</td>
<td>(29.0%)</td>
<td>(25.8%)</td>
</tr>
</tbody>
</table>

**Cattell’s Culture Fair Intelligence Test (CCFIT)**

In order to examine the distribution of intelligence in our sample, I converted the raw CCFIT scores to IQ scores using the norms outlined in the Cattell (1973) manual. The distribution of IQ scores was approximately normally distributed (see Figure 10.2,
with no significant skewness or kurtosis, Shapiro-Wilks $p > .001$) but did contain several low-scoring outliers. A one-sample $t$-test revealed that our sample mean was significantly higher than that of the general population $t(223) = 11.83, p < .001, d = 1.57$. An independent samples $t$-test confirmed that males and females in our sample did not significantly differ in measured intelligence, $t(226) = 1.27, p = .206$. Any observed sex difference in SEI could not, therefore, be explained by apparent differences in actual intelligence resulting from sampling bias. Additionally, a $2 \times (\text{Sex}) \times 4 \times (\text{Sex-Role Category})$ factorial ANOVA confirmed no sex-role differences in measured intelligence, nor any interaction, all $F s < 2.61, p > .05$.

Figure 10.2. Distribution of measured IQ scores in the sample
General and Academic Self-Esteem

I next examined sex and sex-role differences in general self-esteem, using a 2 × (Sex) 4 × (Sex-Role Category) factorial ANOVA (see Figure 10.3). The assumptions of normality and homogeneity of variance were met. There was a significant main effect of sex, $F(1, 219) = 6.71, p = .010, \eta^2 = .03$, with males giving higher self-reports of general self-esteem than females ($d = .40$). Additionally there was a significant main effect of sex-role category, $F(3, 219) = 7.88, p < .001, \eta^2 = .10$, but no interaction between these terms. The effect of sex-role category was stronger than biological sex In line with experimental hypotheses, a planned contrast confirmed that masculine and androgynous subjects reported higher general self-esteem scores than feminine and undifferentiated, $t(225) = 4.62, p < .001, d = .62$, which is a medium effect by Cohen’s (1988) conventions.

Figure 10.3. Rosenberg General Self-Esteem scores across sex and sex-role categories.
Next I examined the construct of academic self-esteem, which was hypothesized as being more tightly coupled to a participant’s self-estimated intelligence score. A $2 \times (\text{Sex}) \times 4 \times (\text{Sex-Role Category})$ factorial ANOVA was conducted on academic self-esteem, and all assumptions were met. As was the case with general self-esteem, males reported significantly higher academic self-esteem than females, $F(1, 219) = 15.01, p < .001, \eta^2 = .06$, as well as a significant main effect of sex-role category, $F(3, 219) = 6.04, p = .001, \eta^2 = .08$. However the interaction was not significant, and again the sex-role identification effect was slightly stronger than biological sex. The planned contrast demonstrated that participants with high masculinity (masculine and androgynous sex roles) reported significantly higher academic self-esteem than participants with low masculinity (feminine and undifferentiated sex roles), $t(225) = 4.26, p < .001, d = .57$, which is a medium effect size.

**Self-Estimated Intelligence (SEI) Scores**

The distribution of self-estimated intelligence scores in our sample was significantly negatively skewed (std. skewness = 2.19), with a tendency for participants to rate their intelligence as “above average”, and a mean SEI of 107.55 ($SD = 10.98$). Surprisingly, quite a number of participants (approximately 19%) rated their intelligence as below average, with scores ranging from 70 IQ points to a maximum of 135. This was unexpected as the ‘above average’ effect is generally robust, and issue I address further in the discussion.

A $2 \times (\text{Sex}) \times 4 \times (\text{Sex-Role Category})$ factorial ANOVA was conducted on self-estimated IQ scores (see Figure 10.4). Although mild negative skewness was present (absolute standardized skewness = 2.23, $p < .05$), the ANOVA is robust against minor

---

4 A reflected log transformation was applied to the distribution and the analysis repeated, with no change in outcome. As the untransformed data was in a metric (IQ score) that was more meaningful, the untransformed data is reported. Additionally the analysis was run with CCFIT as a covariate with no change in outcome.
violations of normality when variances are equal (Field & Wilcox, 2017). The assumption of homogeneity of variance was met. As predicted by prior research, there was a significant main effect of sex, $F(1, 219) = 30.79, p < .001$, $\eta^2 = .12$, with males ($M = 112.12, SD = 9.20$) reporting significantly higher estimated IQ than females ($M = 103.66, SD = 10.88$), $t(225) = 5.55, p < .001, d = .74$, which equates to a difference of approximately 8.5 IQ points. There was also a significant main effect of sex-role category, $F(3, 219) = 7.23, p < .001, \eta^2 = .09$. A planned linear contrast compared the high masculinity participants (masculine + androgynous) to the low masculinity participants (feminine + undifferentiated). As hypothesized masculine and androgynous subjects gave higher self-estimates of IQ than feminine and undifferentiated, $t(225) = 4.65, p < .001, d = .62$. Both effects were medium in size.

There was no significant interaction between sex and sex-role category.

Figure 10.4. Self-estimated IQ scores across sex-role categories, for males and females
Bivariate Correlations

Bivariate correlations between all measures are reported in Table 10.2.

Directions of correlations were consistent with previous literature, with sex and masculinity being significantly associated with self-estimated IQ, both measures of self-esteem, and with IQ discrepancy scores.

Table 10.2

Bivariate Correlations between Sex and Sex-Role Measures, Self-Estimated Intelligence, Measured Intelligence, General and Academic Self-Esteem (N = 228)

<table>
<thead>
<tr>
<th>Measure</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sex</td>
<td>-.21**</td>
<td>.16*</td>
<td>-.38***</td>
<td>-.08</td>
<td>-.20**</td>
<td>-.20**</td>
<td>-.27***</td>
<td></td>
</tr>
<tr>
<td>2. BSRI masculinity</td>
<td>-.02</td>
<td>.34***</td>
<td>.06</td>
<td>.19**</td>
<td>.37***</td>
<td>.26***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. BSRI femininity</td>
<td>-.04</td>
<td>-.07</td>
<td>.04</td>
<td>.11</td>
<td>-.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Self-estimated IQ</td>
<td>-.30***</td>
<td>.44***</td>
<td>.28***</td>
<td>.45***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Cattell IQ</td>
<td>-.72***</td>
<td>-.02</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. IQ Discrepancy</td>
<td>-.22**</td>
<td>.25***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Rosenberg Self-Esteem</td>
<td>-.54***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Academic Self-Esteem</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001

*a Dummy coded variable; 0 = male, 1 = female

Predictors of Sex Differences in Self-Estimated Intelligence

Next I set out to explore possible explanations for the male hubris, female humility effect. In the sample, the correlation between SEI and measured intelligence was just at the cusp of being medium in strength $r(228) = .30, p < .001$, and the scatterplot confirmed it was linear in nature. One possible explanation might be that males and females greatly differ in the accuracy of their judgments of self-estimated intelligence. To test this hypothesis, I calculated the bivariate correlation between SEI
and measured intelligence for males and females separately. The correlation between SEI and measured intelligence was slightly higher for males, \( r(103) = .33, p < .001 \), than for females, \( r(124) = .26, p = .004 \) but both fell in the small to medium range of effect sizes, and any difference most likely reflects sampling error. To confirm this, Fisher’s \( r \)-to-\( z \) transformation was applied to assess the significance of the difference between the two correlation coefficients \( r_{\text{male}} \) and \( r_{\text{female}} \), \( z_{\text{diff}} = .57, p = .284 \) (1-tailed). Thus I was able to rule out glaring differences in accuracy between males and females (see Figure 10.5).

**Figure 10.5.** Scatterplot of association between self-estimated and psychometric IQ, for males and females respectively.

Another plausible explanation for sex differences in SEI might be the contribution of self-esteem. Reported in Table 10.2, there was a moderate positive correlation between self-estimated intelligence and general self-esteem scores. However
it is also plausible that having a high intellect also makes a positive contribution to one’s general self-esteem, so I tested whether the correlation between SEI and general self-esteem remained significant after controlling for scores on the Cattell Culture Fair test. The positive correlation between SEI and Rosenberg General Self Esteem with CCFIT scores partialled out was still statistically significant, \( r = .30, p < .001 \), and of moderate strength (i.e. general self-esteem was associated with self-estimates of intelligence). As might be expected, the correlation between SEI and academic self-esteem was somewhat stronger, \( r = .45 \), though this is likely to be a reciprocal relationship.

To explore the joint effects of biological sex, sex-role identification, and general self-esteem, I performed a hierarchical multiple regression on self-estimated intelligence scores (see Table 10.3). In order to control for individual differences in actual intelligence, at Step 1 I entered Cattell IQ CCFIT scores as the sole predictor, 

\[ F_{\text{chg}}(1,223) = 22.71, p < .001, \] 

explaining approximately 9% of the variance in SEI. Next in Step 2, I entered biological sex, as well as BSRI masculinity and femininity scores. Although only sex and masculinity were hypothesized to make a significant contribution to SEI scores, femininity was included to rule out the possibility of a significant negative association. Together these factors resulted in an increased model fit, 

\[ F_{\text{chg}}(3,220) = 20.76, p < .001, \] 

explaining an additional 20% of variance in the dependent variable. Both sex and masculinity scores were significant predictors. Finally at Step 3, I entered General Self-Esteem scores to test the hypothesis that self-esteem may be a contributing factor. This resulted in a small increase in model fit, 

\[ F_{\text{chg}}(1,219) = 4.39, p < .001. \] 

The final model was statistically significant, \( F(5, 219) = 19.36, p < .001 \), accounting for 31.7% of the variance in individual self-estimates of intelligence. As can be seen from the table, even after controlling for individual differences in measured
intelligence ($\beta = .27$), the three hypothesized predictors of biological sex ($\beta = -.30$), masculinity ($\beta = .21$) and general self-esteem ($\beta = .13$) made significant and unique contributions. Biological sex was the strongest predictor, followed by measured intelligence, masculinity, and finally a smaller contribution of general self-esteem which had considerable overlap with the other predictors.

Table 10.3

Hierarchical multiple regression of Self-Estimated Intelligence scores ($N = 255$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$-value</th>
<th>$sr^2$</th>
<th>$R$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cattell IQ</td>
<td>.30</td>
<td>4.77</td>
<td>&lt;.001***</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td>.54</td>
<td></td>
<td></td>
<td>.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cattell IQ</td>
<td>.26</td>
<td>4.61</td>
<td>&lt;.001***</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (0 = male)</td>
<td>-.32</td>
<td>-5.46</td>
<td>&lt;.001***</td>
<td>.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masculinity</td>
<td>.26</td>
<td>4.42</td>
<td>&lt;.001***</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Femininity</td>
<td>.03</td>
<td>.58</td>
<td>.562</td>
<td>.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Step 3</strong></td>
<td>.55</td>
<td></td>
<td></td>
<td>.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cattell IQ</td>
<td>.27</td>
<td>4.72</td>
<td>&lt;.001***</td>
<td>.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (0 = male)</td>
<td>-.30</td>
<td>-5.09</td>
<td>&lt;.001***</td>
<td>.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masculinity</td>
<td>.21</td>
<td>3.44</td>
<td>.001**</td>
<td>.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Femininity</td>
<td>.02</td>
<td>.28</td>
<td>.780</td>
<td>.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Self-Esteem</td>
<td>.13</td>
<td>2.10</td>
<td>.037*</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$, *** $p < .001$

Next I examined whether masculine sex-role identification (masculinity score as a continuous variable) acted as a statistical mediator in the relationship between biological sex and SEI scores (see Figure 10.6) to test hypothesis 5. Baron and Kenny (1986) proposed three criteria for establishing statistical mediation. Firstly the predictor (biological sex) should predict the dependent variable (SEI). Secondly, the predictor must be correlated with the proposed mediator variable (masculine sex-role
identification, shown as Path A). Thirdly the mediator must correlate with the dependent variable (SEI) even after controlling for the contribution of the predictor (shown as Path B). The Sobel test of statistical mediation was significant, Sobel $z = -2.55$, $p = .010$, and calculation of the bootstrapped estimate of the indirect effect showed that it differed significantly from zero [95% CI = -2.26 to -0.41], following the criteria outlined in Preacher and Hayes (2004). As the mediation effect was significant, I then tested whether the relationship was fully or only partially mediated (Baron & Kenny, 1986). In a full mediation model, the association between predictor and dependent variable will no longer be statistically significant after controlling for the mediator (i.e., all of the effect of the predictor acts indirectly through the mediator, and does not make a direct contribution). This relationship is represented by Path C in Figure 10.6. Though diminished, the beta weight remained statistically significant, indicating that the relationship was only a partial mediation. Though acting indirectly through masculine sex-role identification, there was still a direct contribution of sex to SEI scores.

**Figure 10.6.** Indirect effect of sex on SEI, with masculine sex-roles acting as a mediator on self-estimated intelligence. Path C represents the direct effect of sex after controlling for the mediator.

Having identified in the multiple regression analysis that biological sex made a slightly stronger contribution to SEI than measured intelligence, sex-role identification,
and general self-esteem, I sought to quantify how large the discrepancy between self-estimates and measured intelligence was. A composite variable representing the discrepancy between self-estimated and measured intelligence was created, with positive values indicating higher SEI than measured intelligence. An independent samples \( t \)-test on IQ discrepancy scores confirmed a significant sex difference, \( t(225) = 3.00, p = .003, d = .40 \). Visual inspection of the discrepancy scores showed that on average, males in our sample demonstrated fairly sound judgement in appraising their intelligence (\( M = -0.36, SD = 13.60 \)), but that there was also wide variability with some males greatly overestimating their intelligence and some males underestimating (range = -27 to +38 IQ points). However females systematically undervalued their intellectual capabilities by over 6 IQ points (\( M = -6.32, SD = 15.89 \)), and for those female participants who did offer inflated self-estimates, these were much smaller in size (range = -41 to +25 IQ points).

To confirm our interpretation of the data, a categorical variable named accuracy direction was created to measure whether participants had underestimated their intelligence (discrepancy score < -5 IQ points), overestimated their intelligence (discrepancy score > +5 points), or made an accurate assessment (in the range -5 to +5 IQ points discrepancy). Though arbitrary, this represents a discrepancy between perceived and objectively assessed IQ of one third of a standard deviation or Cohen’s \( d = .33 \). Chi-square analysis (see Table 10.4) showed that there were significant sex differences in accuracy direction, \( \chi^2 (2, N = 228) = 7.26, p = .027 \). Inspection of the adjusted standardized residuals showed that there were significantly more females underestimating their intelligence than males \( (adj. z = 2.7, p = .004 \) 1-tailed). In particular, over half of the female participants significantly underestimated their intellectual ability compared to only a third of the male participants. Consistent with the
male hubris effect, there were also significantly more males than females overestimating their intelligence \((adj \ z = 1.8, p = .035)\) 1-tailed). However, if a non-directional 2-tailed comparison were used, the sex difference in overestimators would just fall short of statistical significance.

Table 10.4

**Accuracy direction of male and female participants**

<table>
<thead>
<tr>
<th>Accuracy direction</th>
<th>Sex</th>
<th>Underestimating</th>
<th>Accurate</th>
<th>Overestimating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>35.9%</td>
<td>32.0%</td>
<td>32.0%</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>53.6%</td>
<td>24.8%</td>
<td>21.6%</td>
<td></td>
</tr>
</tbody>
</table>

**Self-estimates of multiple intelligences**

Next, a \(2 \times \text{Sex} \) \(4 \times \text{Sex-Role Category}\) factorial MANOVA was performed on the nine self-estimates of Gardner’s multiple intelligences. As the cell size differed across sex-role category and Box’s \(M\) was significant \((p < .001)\), Pillai’s trace was selected as the more conservative estimate. Assumptions of normality and homogeneity of variance were met. In line with previous research, there was a significant multivariate effect of biological sex, \(F(9, 212) = 7.02, p < .001, \eta^2 = .23\) which is a medium to large effect. There was also a significant multivariate effect of sex-role identification, \(F(27, 642) = 2.22, p < .001, \eta^2 = .09\), though there was no significant interaction \(F(27, 642) = 1.02, p = .437\). As the overall multivariate effects were significant and of non-trivial size, this justified examination of univariate effects without a need to apply a Bonferroni correction (c.f. Huberty & Morris, 1989). For ease of comparison, sex and sex-role differences are reported separately in Table 10.5 and 10.6 respectively. Five of
the nine multiple intelligence domains showed significant differences between males and females, with effect sizes ranging from small to large.

Table 10.5

*Sex differences on self-estimated multiple intelligences*

<table>
<thead>
<tr>
<th>Domain</th>
<th>Male</th>
<th>Female</th>
<th>$F_{(1,220)}$</th>
<th>$p$-value</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verbal</td>
<td>106.45 (12.87)</td>
<td>107.07 (11.65)</td>
<td>.73</td>
<td>.395</td>
<td>-.05</td>
</tr>
<tr>
<td>2. Logical-Mathematical</td>
<td>108.39 (16.68)</td>
<td>98.66 (13.43)</td>
<td>18.36</td>
<td>&lt;.001***</td>
<td>.64</td>
</tr>
<tr>
<td>3. Spatial</td>
<td>109.80 (12.51)</td>
<td>98.54 (11.93)</td>
<td>40.79</td>
<td>&lt;.001***</td>
<td>.92</td>
</tr>
<tr>
<td>4. Musical</td>
<td>102.64 (18.11)</td>
<td>99.50 (14.72)</td>
<td>.64</td>
<td>.426</td>
<td>.19</td>
</tr>
<tr>
<td>5. Bodily-kinaesthetic</td>
<td>112.57 (14.26)</td>
<td>106.47 (14.74)</td>
<td>7.54</td>
<td>.007**</td>
<td>.42</td>
</tr>
<tr>
<td>6. Interpersonal</td>
<td>112.69 (12.98)</td>
<td>112.86 (11.72)</td>
<td>.20</td>
<td>.654</td>
<td>-.01</td>
</tr>
<tr>
<td>7. Intrapersonal</td>
<td>110.61 (12.63)</td>
<td>109.36 (12.79)</td>
<td>.11</td>
<td>.742</td>
<td>.09</td>
</tr>
<tr>
<td>8. Naturalistic</td>
<td>104.43 (11.88)</td>
<td>99.10 (11.06)</td>
<td>10.36</td>
<td>.001**</td>
<td>.46</td>
</tr>
<tr>
<td>9. Existential/spiritual</td>
<td>108.72 (16.92)</td>
<td>102.94 (12.84)</td>
<td>6.85</td>
<td>.009**</td>
<td>.39</td>
</tr>
</tbody>
</table>

* $p < .05$; ** $p < .01$; *** $p < .001$;

Table 10. presents sex-role differences across the nine multiple intelligence domains. Although sex differences were not present for every domain (Table 10.4), there were significant sex-role differences for each of the domains. Accordingly a planned linear contrast was conducted comparing the high masculinity groups (masculine + androgynous) with the low masculinity groups. Masculine persons reported significantly higher self-estimates of multiple intelligences, with effect sizes ranging from small to medium in size.
Table 10.6

Sex-role Differences in Self-Estimated Multiple Intelligences

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verbal</td>
<td>110.35</td>
<td>104.51</td>
<td>106.74</td>
<td>104.54</td>
<td>2.78*</td>
<td>t(226) = 2.44, p = .015, d = .33</td>
</tr>
<tr>
<td>F-ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.78</td>
<td>t(226) = 2.32, p = .021, d = .32</td>
</tr>
<tr>
<td>2. Logical-Mathematical</td>
<td>106.66</td>
<td>98.41</td>
<td>104.38</td>
<td>103.23</td>
<td>2.68*</td>
<td>t(226) = 3.22, p = .001, d = .43</td>
</tr>
<tr>
<td>(1.98)</td>
<td></td>
<td></td>
<td>1.85</td>
<td>2.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Spatial</td>
<td>105.71</td>
<td>100.34</td>
<td>107.01</td>
<td>101.83</td>
<td>3.70*</td>
<td>t(226) = 4.18, p &lt; .001, d = .56</td>
</tr>
<tr>
<td>(1.59)</td>
<td></td>
<td></td>
<td>1.49</td>
<td>1.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Musical</td>
<td>102.98</td>
<td>95.86</td>
<td>105.23</td>
<td>97.40</td>
<td>4.25**</td>
<td>t(226) = 4.40, p = .003, d = .40</td>
</tr>
<tr>
<td>(2.12)</td>
<td></td>
<td></td>
<td>1.99</td>
<td>2.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Bodily-kinaesthetic</td>
<td>112.18</td>
<td>106.22</td>
<td>114.00</td>
<td>103.96</td>
<td>6.47***</td>
<td>t(226) = 4.18, p &lt; .001, d = .56</td>
</tr>
<tr>
<td>(1.86)</td>
<td></td>
<td></td>
<td>1.75</td>
<td>1.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Interpersonal</td>
<td>113.82</td>
<td>112.36</td>
<td>117.87</td>
<td>105.83</td>
<td>10.82***</td>
<td>t(226) = 4.30, p &lt; .001, d = .57</td>
</tr>
<tr>
<td>(1.53)</td>
<td></td>
<td></td>
<td>1.44</td>
<td>1.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Intrapersonal</td>
<td>110.48</td>
<td>107.58</td>
<td>113.88</td>
<td>106.44</td>
<td>4.09**</td>
<td>t(226) = 3.03, p = .003, d = .40</td>
</tr>
<tr>
<td>(1.66)</td>
<td></td>
<td></td>
<td>1.56</td>
<td>1.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Naturalistic</td>
<td>101.83</td>
<td>100.49</td>
<td>104.20</td>
<td>99.12</td>
<td>2.15</td>
<td>t(226) = 2.09, p = .038, d = .27</td>
</tr>
<tr>
<td>(1.50)</td>
<td></td>
<td></td>
<td>1.41</td>
<td>1.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Existential/spiritual</td>
<td>106.80</td>
<td>104.23</td>
<td>110.47</td>
<td>100.57</td>
<td>4.79**</td>
<td>t(226) = 3.15, p = .002, d = .42</td>
</tr>
<tr>
<td>(1.93)</td>
<td></td>
<td></td>
<td>1.81</td>
<td>1.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p < .05; ** p < .01; *** p < .001;

Discussion

The main aim of this study was to investigate factors that may explain why males tend to provide higher self-estimates of their intelligence than females. A meta-analysis by Szymanowicz and Furnham (2011) demonstrated that the male-hubris, female humility effect was robust, with an average effect size of $d = .37$, but that this was somewhat higher in samples drawn from psychology and social science subject pools, $d = .48$. Furham and Rawles (1995) have attributed this effect to a misbelief held by psychology students about the intellectual superiority of males. Still, sex differences in SEI for our sample ($d = .74$) were somewhat higher in our sample than would be
expected based on previous studies with psychology subject pools (Szymanowicz & Furnham, 2011), falling into the medium to large range. Importantly I was also able to rule out the possibility of an actual sex difference in measured intelligence in our sample (arising due to sampling bias from a university educated subject pool). Group differences between males and females in SEI were illusory and did not reflect actual intelligence differences. Consistent with hypotheses there was also an independent effect of sex-role identification in SEI, with masculine and androgynous subjects reporting higher self-estimates than feminine and undifferentiated, $d = .62$ which is also medium to large effect size.

**Explanations for Sex Differences in Self-Estimated Intelligence**

An important research question has been whether people hold realistic views of their own intelligence, or whether they are distorted. This question has been of broader interest to researchers (irrespective of gender), with a large number of studies investigating whether people are accurate judges of their own ability (Kruger & Dunning, 1999). The relationship between self-estimated and psychometrically measured intelligence is generally fairly weak (Paulhus et al., 1998), suggesting that while actual ability does contribute to the self-image that we hold, there are other factors at play.

A limitation of the studies mentioned above is that they did not test whether there were sex differences in accuracy of intellectual perceptions (for example, if males held somewhat inflated perceptions while females were more pragmatic). Indeed relatively few studies have explicitly tested the hypothesis that there might be dramatic sex differences in accuracy of perceptions. Some studies have found weaker correlations in females than males for the relationship between SEI and measured intelligence, while other studies have found their accuracy to be comparable. For example, Borkenau and
Liebler (1993) found the correlation between SEI and measured intelligence to be equivalent for males and females. Similarly, Reilly and Mulhern (1995) found similar correlations between SEI and measured intelligence for males and females. However, a third study by Furnham and Rawles (1999) found that only males showed a significant correlation between SEI and measured intelligence. Another study by Furnham and Fong (2000) using the Raven’s Progressive Matrices as a measure of fluid intelligence also provided mixed findings in a sample of British and Singaporean students. In the British sample, the relationship between SEI and measured intelligence was significant for males but not for females. However, both males and females showed a similar degree of accuracy in the Singaporean sample with a medium effect size.

In our sample, the correlation between SEI and measured intelligence was moderate in strength, \( r = .30 \). Furthermore, there was a significant association for both males and females and no significant difference in correlation strengths (see Figure 10.5). Thus it is possible to rule out gross disparities in accuracy between males and females as one possible explanation (though there were still discrepancies in direction, with a self-enhancing bias in males and self-deriding bias in females). Next, I examined the discrepancy scores and accuracy direction. Consistent with the male hubris, female humility effect, chi-square analysis showed that there was a greater frequency of males overestimating their intelligence than for females. Furthermore there were also significantly more females than males underestimating their ability.

One reason for these directional differences this might be sex differences in modesty norms (Rudman, 1998), and social pressure on women to demonstrate public modesty. But given that sex differences in intellectual self-concept are seen so early in childhood, another plausible explanation for sex differences in SEI might be the

---

5 No relation to the present author
contribution of self-esteem, explored by Hypotheses 3 and 4. Although most people tend to see themselves in a rosy light, as more honest, more attractive, more intelligent than the average person (Kruger & Dunning, 1999), for people with low general self-esteem, views of the self are decidedly negative, especially when making social comparisons to others. Consistent with Gabriel, Critelli and Ee (1994), self-estimated intelligence and general self-esteem were related with a moderately sized correlation. Having a positive self-image of oneself generally leads to higher self-reports of intelligence with a similarly sized contribution as psychometrically measured intelligence. The analysis also found the discrepancy between SEI and measured SEI was positively associated with self-esteem (i.e. people with high self-esteem showed a tendency to overestimate their intelligence). Given the widely documented sex differences in self-esteem (Kling et al., 1999), as well as literature on the contribution of sex-role identification to self-esteem (Hirschy & Morris, 2002; Whitley & Gridley, 1993) and in light of the present findings, self-esteem warrants further investigation in future SEI studies as one mechanism underlying sex differences in SEI. It also suggests that sex differences in SEI are genuinely held beliefs, countering the argument that they might be due to modesty norms and differ from actual self-concept (Rudman, 1998).

Next I used a hierarchical regression model to determine the joint effects of sex, sex-role identification, and general self-esteem. Each made a significant and unique contribution to overall SEI scores, even after statistically controlling for actual fluid intelligence. Masculinity was positively correlated with higher SEI, but there was no effect of femininity as a personality trait. The relationship between sex and SEI was partially mediated by masculine sex-role identification, and met the criteria for statistical mediation. Thus there was an indirect effect of sex on SEI scores mediated through masculinity, as well as a direct effect. Much of the contribution of self-esteem
overlapped with sex and sex-role identification (as would be expected given the
documented sex/sex-role effects in the literature) but it still made a significant and
unique contribution.

**Self-estimates of multiple intelligences**

Next, I examined sex and sex-role differences in estimates of multiple intelligences. Only five of the nine multiple intelligences showed statistically significant sex differences. Consistent with the meta-analysis by Szymanowics and Furnham (2011), logical-mathematical and spatial showed the largest differences, but meaningful effect sizes were also found for bodily-kinaesthetic, naturalistic and existential/spiritual intelligence. Visser, Ashton and Vernan (2008) have also reported a moderately large sex difference on self-estimates of bodily-kinaesthetic intelligence, but relatively few studies have evaluated naturalistic and existential multiple intelligences as they were only recently added to Gardner’s revised model. However Furnham and Ward (2001) reported significantly higher estimates by males for naturalistic intelligence, and a study by Yuen and Furnham (2006) found significant non-trivial sex differences for bodily-kinaesthetic, naturalistic, and existential/spiritual intelligence consistent with our pattern of results. Intelligence domains that did not show significant differences in self-estimates were verbal, musical, and the two aspects of emotional intelligence (intrapersonal and interpersonal). Perhaps the reason is that gender stereotypes for these domains generally favour women (Bennett, 2000), thus offsetting the tendency towards higher estimation by for intelligence generally. It might also be the case that women are less likely to underestimate their abilities here and more likely to underestimate in domains associated with males, given popular gender stereotypes. A number of previous studies have also found non-significant effects for these domains (Furnham, Clark, & Bailey, 1999; Rammstedt & Rammsayer, 2002b), which would be consistent with the
interpretation that gender stereotypes favouring women buffers against the trend towards higher male self-estimates generally.

While sex differences were found for some, but not all, multiple intelligences domains, the questionnaire was more sensitive to sex-role effects. Planned contrasts revealed that the high masculinity groups (masculine + androgynous) reported significantly higher self-estimates than the low masculinity groups (feminine + undifferentiated), regardless of biological sex. Just as with general self-esteem and academic self-esteem, masculinity seemed to have an enhancing effect on self-estimates of intelligence in these domains. Several other studies have investigated sex-role identification effects on the estimates of multiple intelligences with mixed findings. Rammstedt and Rammsayer (2002a) used the BSRI in a German sample, finding sex-role identification acted as a moderator for some but not all domains. Furnham, Clark and Bailey (1999; Study 2) recruited a small sample of students ($n = 80$) using the PAQ instrument to measure sex-role identification and did not find sex-role effects, but their study was extremely underpowered. More recently, Szymanowicz and Furnham (2013) recruited a British sample from the general population, finding significant sex-role effects for verbal, social, emotional, and practical intelligence factors, but not the expected effect for mathematical/logical and spatial intelligence. However subsequent regression analyses that treated masculinity as a continuous variable showed a significant positive association on all but the emotional intelligence factor. Thus our study replicates their findings that masculinity leads to higher self-estimated intelligence scores for multiple intelligence domains.

**Implications and limitations**

Our study expands on the existing body of literature on self-estimated intelligence by ruling out dramatic sex differences in accuracy, but also showing that
general self-esteem is as strongly correlated with SEI as is psychometrically measured intelligence. Having a positive self-image is correlated with higher self-estimated intelligence, though causality cannot be determined. Significant sex and sex-role differences in self-estimated intelligence were found despite there being no difference in psychometrically measured intelligence. In particular, masculinity acted as a protective factor while there was no effect of femininity. The relationship between sex and self-estimated intelligence was partially mediated by masculine sex-role identification, but there remained a significant direct effect of sex as well. Additionally, significant sex-role differences were found on all nine multiple intelligence domains, even in the absence of sex differences.

The effect we were investigating has been termed by Furnham et.al (2001) as the “male-hubris, female humility” effect. This phraseology stems from the observation that males rate their intellectual abilities as being much higher than females, despite compelling psychometric evidence that sex differences in general intelligence do not exist (Halpern, 2011; Jensen, 1998). In our study, the inclusion of the Cattell Culture Fair Intelligence Test afforded the opportunity to examine the discrepancy between SEI and measured IQ, and the nature of sex differences. Although significant sex differences in SEI were observed, especially for direction of the discrepancy, as noted earlier the average discrepancy score in IQ for males approached zero with inflated scores being largely offset by under-estimates. This finding seems peculiar, and requires further investigation. It may have one of several explanations. It may be a property of the sample recruited in the present study: our sample consisted of primarily first-year university students, with a mixture of below-average IQ (15.5%) and mature-age students who may be lacking in self-confidence and doubting their intellectual ability.
academic self-esteem (as well as a significant correlation between academic self-esteem and discrepancy scores), so our sample may differ from other student samples who have been studying longer and are more settled. Alternately, it may be that college subject pools drawn from American samples have higher self-esteem, as has been claimed by some researchers (Baumeister, Campbell, Krueger, & Vohs, 2005; Diener & Diener, 1995). Another explanation may be an unintended consequence of the informed consent and briefing materials, in that participants were aware they would soon complete an intelligence test to compare against their self-estimates. This knowledge might have dampened self-estimated IQ scores somewhat, leaving some males to provide more conservative estimates than they would otherwise without the knowledge that their scores would be actually checked. However, as demonstrated by the chi-square analysis of accuracy of estimates there were still a large number of males with inflated self-estimates, as much as 38 IQ points – so if it did exert an effect it did not do so uniformly.

Another possibility is that it might be a property of the Cattell Culture Fair Test itself. Although an excellent measure of fluid intelligence, the instrument has not been revised since 1973, and the provided norms tables may have resulted in inflated IQ scores. The well documented Flynn effect of rising IQ scores across historical time may have been an issue. A recent meta-analysis by Pietschnig and Voracek (2015) has found that rises in intelligence vary depending on the domain, with fluid intelligence (measured by the Cattell instrument) showing the greatest gains, approximately 0.41 per annum. While the instrument remains a reliable and valid measure of nonverbal fluid intelligence to include in a regression of self-estimated intelligence, the outdated norms tables may have yielded an inflated IQ score for calculating discrepancy scores. Thus I cannot rule out the possibility that, on average, males are still overestimating their level
of intelligence in line with the male-hubris effect, but that the outdated norm table resulted in a discrepancy score in our sample for males approaching zero. Nonetheless, I did see a substantial underestimation for the females in our sample, consistent with the female humility effect. So few studies investigate discrepancy scores, and it would be interesting to see if there is still a pattern ‘male hubris’ when intelligence is measured by other instruments that provide a Full-Scale IQ such as the WAIS or Stanford-Binet tests.

**Conclusions**

How we see ourselves intellectually can have a profound influence on academic engagement. Self-perceptions that are completely out of touch with reality are problematic: when unrealistically high it may set students up for eventual failure and disappointment, but when unrealistically low it may result in student disengagement, lowered academic expectations, and a failure to pursue educational options that a student is capable of achieving. Research has identified that academic self-concept is strongly linked to academic achievement, and that this is a reciprocal relationship (Marsh, 1990; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Valentine, DuBois, & Cooper, 2004) While the pattern of higher self-estimates of intelligence by males than females is a robust effect replicated cross-culturally, relatively little progress had been made in understanding why this effect exists. Our results point to two contributing factors, firstly the influence of general self-esteem and secondly sex-role identification. Issues of low self-esteem are generally considered in the context of mental health and psychosocial wellbeing, but our results show that it may partly contribute to the so-called ‘male hubris, female humility’ effect as well. Starting from adolescence onwards, girls and women are overrepresented in depressive symptoms even in non-clinical populations (Wang et al., 2016), and lowered self-esteem is one consequence of this.
Our results suggest that lowered self-esteem underlies the male hubris, female humility effect, and is an underappreciated risk factor for educational achievement. Though dated, a meta-analysis by Hansford and Hattie (1982) reported a positive association between self-esteem and academic achievement \((r = .22)\), supporting such a conclusion that a negative intellectual self-concept may translate to poorer educational outcomes. Eccles (1994, 2013) expectancy-value model of achievement motivation suggests that a crucial factor in achievement-related choices is their perceptions of their own abilities and expectations of success.

Additionally a large body of research (e.g. Else-Quest, Mineo, & Higgins, 2013) has reported lower female self-efficacy beliefs in mathematics and science, and suggested that these may be far more influential than actual ability in contributing to the underrepresentation of women in STEM. However these have almost always been interpreted in light of cultural stereotypes associating STEM with masculinity (Nosek et al., 2002), rather than considering the contribution of a negative intellectual self-image more broadly. In planning educational interventions to raise STEM performance, addressing academic self-esteem and intellectual self-image may be an important targets for consideration. Some educational interventions such as values affirmation writing exercises (Kost-Smith et al., 2012; Miyake et al., 2010) buffer against intellectual stereotype threat, and may be particularly useful in an educational setting in combatting lowered intellectual self-image (Martens, Johns, Greenberg, & Schimel, 2006). Also while sex-role identification is not a desirable target for intervention, it may identify students of either gender prone to underestimating their intellectual abilities, whom parents and educators can be mindful of as needing further support and encouragement.
References


Chapter 11 - Discussion

The time has come, ’ the Walrus said,
To talk of many things:
Of shoes — and ships — and sealing-wax —
Of cabbages — and kings…. Lewis Carroll: ‘The Walrus and the Carpenter’

This chapter serves as an overview and integration of the collected studies reported in the thesis, and how they have addressed the four research questions outlined in Chapter 1. The overarching goal of this program of research was to make progress on a seemingly intractable problem – why do sex differences develop for specific cognitive abilities at the population level, and are there alternative explanations (such as psychological traits, self-concept) for the group differences? However, there was a lack of consensus in the literature about whether sex differences still exist, and if so to what extent are they present in the population.

The current thesis had two major aims. The first aim was to address identified gaps in the literature about the existence and magnitude of sex differences with contemporary samples, and to provide a firmer evidence base by using representative samples and cross-cultural data-sources. Furthermore, it has been argued that cross-cultural studies in particular offer the opportunity to test various theories about origins – if sex differences are universal and show minimal variability, then it would at least be consistent with a biological cause (Geary, 2010; Kenrick, Trost, & Sundie, 2004). If they were universal but showed meaningful variability, it would suggest that there are social and cultural practices that act as moderators. And if they were quite inconsistent (either reversing direction, or varying greatly in magnitude), this pattern would support psychosocial models but contradict claims of innate and immutable biological differences.
The second aim was to address the more difficult problem of why sex differences are found, by exploring a variety of psychosocial theories: including Nash’s sex-role mediation theory, the effect of situational factors like the testing environment on performance (even in the absence of high stakes testing), and the contribution of self-estimated intelligence. Using the information gleaned from studies addressing the second aim, a refinement of existing psychobiosocial models is proposed, situating distal and proximal factors. Each research question is addressed in turn.

11.1 Magnitude of sex differences in cognitive abilities

Much of the research on sex differences in cognitive abilities at the time of starting this research program was dated, and there was the very legitimate question of whether the research findings of the past would generalise to modern samples growing up in more egalitarian times (didn’t we solve that whole ‘gender’ thing, right?). Feingold (1988) made the bold claim that cognitive differences were disappearing, while Hyde (2005) advanced the ‘gender similarities hypothesis’ which holds that most sex differences are actually small or trivial in magnitude and that future research should be framed in terms of gender similarities rather than gender differences. Based largely on these two lines of evidence, Caplan and Caplan (2016) questioned whether sex differences in verbal and language abilities existed, and suggested that the motives for conducting further research were alpha bias or ideological in nature (Caplan & Caplan, 1997).

The gender similarities hypothesis in particular has exerted a strong influence on subsequent literature in the field, and has to some degree constrained further scientific research in this area (Eagly, 2018; Halpern, 2014b). But Hyde and Grabe (2008) also made a compelling argument that the technique of meta-analysis (allowing one to investigate effects over a range of studies, sample types, and time-points) can offer
greater clarity in understanding the extent of sex differences or their absence. As Rosenthal (1995) noted, meta-analytic reviews have much more to offer than merely attesting to the size and robustness of an effect: they also afford the ability to examine potential moderators such as developmental effects, and historical changes over time. This is especially important in the field of sex difference research, as there is often sampling bias due to non-representative convenience samples (Becker & Hedges, 1988) which can distort the conclusions drawn (Wilkinson, 1999).

Take, for example, the issue of mathematical and scientific abilities (collectively referred to as quantitative reasoning). In a high profile study published in *Science*, Hyde, Lindberg, Linn, Ellis, and Williams (2008) analysed state performance data collected in the United States for the NAEP. They reported that the weighted mean sex difference across ages was $d = .0065$ and essentially trivial in every grade. A limitation of their methodology was that it was a convenience sample limited to only ten states, and was a snapshot in time across a single year (unstated). Furthermore, it only examined mean sex differences in mathematics rather than sex ratios at the tail of the distribution, as reported by Hedges and Nowell (1995). Curiously, though, the paper claimed gender similarities in mathematics and science without reporting analysis of any science achievement (but is arguably germane to the issue of STEM). There are various reasons why Hyde et al.’s data might not be representative of the nation as a whole (state assessment data often draws only from public schools not private, and there can be disparities in the educational standards and curricula not only at the local county level in the United States, but also across states). However, public availability of national testing data from NAEP assessments offered the opportunity to hold such a claim to greater scrutiny (see Chapter 4).
Chapter 4, reported as Reilly, Neumann and Andrews (2015) revealed a more complex and more nuanced picture than the conclusion of gender similarities found by Hyde et al. (2008). Rather than an absolute magnitude of $d = .00$ reported by Hyde et al., a small but stable sex difference in mathematics was observed with a developmental trend towards a peak after adolescence in Grade 12 of around $d = +.10$. It pointed to neither a substantial gap nor a trivial amount, especially when investigating sex ratios of high achieving students attaining the ‘Advanced’ proficiency standard in maths. By the end of compulsory schooling, the sex ratio was over twice as many males as females (2.13) for mathematics – certainly a cause for concern and target for further study. Importantly, there was no evidence for a decline in either the effect size or the sex ratio over time for the period analysed (1990 - 2011) as had been claimed by Feingold (1988) and Caplan and Caplan (2005).

Furthermore, Reilly et al. (2015) conducted the first analysis of science achievement data from the NAEP since Hedges and Nowell’s (1995) pioneering review over thirty years ago. Sex differences in science achievement were also relatively modest, $d = +.11$, but showed a similar developmental trend towards larger differences in older students. This effect varied by science discipline, with somewhat larger effects found in earth and space sciences ($d = +.21$ by Grade 12) and physical sciences ($d = +.18$ by Grade 12), and importantly no significant difference for biology and life sciences - even given the appreciable sample size. These findings point not to any inherent lack of ability (not that this was ever predicted!), but rather potential differences in interest level and relevance to the type of careers. A comprehensive meta-analysis by Su, Rounds and Armstrong (2009) of over half a million respondents showed that men have greater interests in things, and women show greater interest in people, with a large sex difference on the Things-People dimension ($d = 0.93$).
Interestingly medicine and biological sciences are the only STEM fields where equal or greater representation of women is reached (National Science Foundation, 2017).

Furthermore, by Grade 12 there were over twice as many males than females achieving the ‘Advanced’ proficiency standard set by the NAEP (2.28), which has bearing on the relative numbers of men and women seeking to pursue STEM careers. Voracek, Mohr, and Hagmann (2013) have argued the importance of considering these tail ratios, especially in the context of sex difference research but notes that they are scarcely investigated.

My purpose in restating these findings is to observe that it revealed a very different picture than had been previously offered by Hyde and colleagues (which essentially amounted to endorsement of the null hypothesis). By asking the unasked research questions (are there sex differences in high-achieving students, are there developmental effects) new information was revealed. We also extended research to consider the question of sex differences in science achievement, and important moderators (developmental, and by scientific field) that had been overlooked in prior research.

There is an expression in the legal profession termed the ‘chilling effect’ that restrictions on freedom of speech or the threat of a lawsuit can have on subsequent public discourse. Meta-analysis is imbued with a special power and respect in psychology and the social sciences (when in doubt, consult a meta-analysis!) and so carries greater evidentiary weight. But as Rosenthal and DiMatteo (2001) observed, their power to shed light on research questions are limited by the quality of the data used and the research questions asked. I would argue that a similarly ‘chilling effect’ can be present for subsequent scientific research when a highly visible, compellingly written meta-analytic review concludes support for the null hypothesis and that the
debate is now closed. The pattern of Hyde’s analyses were also inconsistent with international studies of mathematics (Guiso, Monte, Sapienza, & Zingales, 2008; Chapter 6), and so seemed ‘anomalous’. Becker and Hedges (1988) long ago argued the importance of considering the effects of selection bias in recruitment of samples when testing for sex differences, due to a variety of factors including interactions with demographic variables such as socioeconomic status (Hanscombe et al., 2012; Turkheimer, Haley, Waldron, D’Onofrio, & Gottesman, 2003), rural versus regional localities, and ethnicity (Else-Quest, Mineo, & Higgins, 2013).

Another domain of cognitive abilities where fairly firm support for the existence of sex differences had been found by Maccoby and Jacklin (1974) was that of verbal and language abilities. Indeed, the existence of sex differences in verbal abilities had ‘been one of the tried and true “facts” of psychology for decades’ (Hyde & Linn, 1988, p. 53). Yet in the meta-analysis conducted by Hyde and Linn, they had concluded that year of publication was an important and previously overlooked moderator, such that sex differences in verbal abilities were decreasing overtime. In studies published in 1973 or earlier, the effect size was $d = -0.23$ – but in studies published after 1973 the effect size was considerably smaller $d = -0.10$. On this basis, Hyde and Linn concluded that “the difference is so small that we argue that gender differences in verbal ability no longer exist”, (p. 53). I have addressed the shortcomings of the Hyde and Linn meta-analysis in Chapter 2 (succinctly a cherry-picking of literature which Stumpf (1995) reviews) and how their conclusions differ starkly from other sex difference researchers (Halpern, 2000, 2011; Kimura, 2000). Yet it is not uncommon to still find it cited as conclusive evidence that sex differences in verbal abilities do not exist, and it formed the basis for Caplan and Caplan’s (1997) assertion that sex differences in verbal and language abilities had been eliminated.
This position was refuted by later reviews (such as Hedges & Nowell, 1995), but it did ask an intriguing question – if not absent, were they at least *diminishing* in response to changes in societal values or educational practices? Though many waves of NAEP assessment data had been collected in the decades since Hedge and Nowell’s review, it came as genuine surprise to me that there were no published analyses on the dataset examining the extent of sex differences in reading and writing. As before with the meta-analysis of mathematics and science achievement, it struck me as an unasked research question (or if it had been asked and analysed, presumably gone unpublished\(^6\)).

These claims though by two prominent sets of researchers now meant that there was a lack of consensus in the literature (verbal differences were either “well established” or entirely spurious). To their credit, Caplan and Caplan (1997) did raise a genuine concern about methodological issues with studies, such as employing non-representative convenience samples, and the vagaries of operational definitions of verbal ability. Reading and writing proficiency were at least clearly well defined, and the NAEP provided a vigorous assessment framework that remained stable over time.

Chapter 5, and subsequently published as Reilly, Neumann and Andrews (2018) put this research question to the test. Our study found compelling evidence of mean sex differences in reading and writing, as well as in the sex-ratios of students attaining the lowest and highest proficiency standards for both outcomes. There were also developmental effects towards larger differences as students progress through schooling, and contrary to Feingold’s claim there was no decline in magnitude over the timespan investigated (27 years). I would argue that the difference from Hyde and Linn

---

\(^6\) During extensive rounds of peer review, a concern echoed by several reviewers was that there were danger in reporting sex differences, for fear that it might be misused by lay advocates of single-sex schooling or to support claims of biological determinism. The corollary to omitting them from the publication record is that it stifles research into their aetiology and educational interventions to reduce the size of the gender gap
(1988) was selective inclusion/exclusion of certain types of verbal ability that showed larger differences (e.g., verbal fluency, writing, spelling, grammar and language usage), and the non-representative nature of their samples. Our findings concur with another study published in the same month by Petersen (2018) on state-based NAEP assessments, which found that sex differences in verbal ability are robust and generalise to tasks other than just reading.

As mentioned above, cross-cultural research offers a stronger evidence base than that drawn from a single country. So in Chapter 6, I examined cross-cultural patterns of sex differences in reading, mathematics and science achievement. For reading, all countries investigated showed significantly higher female performance (consistent with a biological contribution), but interestingly there was also substantial variability in the magnitude across countries, an effect first observed by Guiso et al. (2008). Countries with greater gender equality showed larger sex differences in reading achievement, and this observation has been replicated in all existing waves of PISA assessment (Reilly, 2015). Global sex differences were also found for mathematics, though the effect was stronger in OECD nations and correlated with national levels of gender equality and a country’s tolerance for wealth inequality. A somewhat different pattern of sex differences was found for science achievement, with some countries showing substantial sex differences favouring males and others favouring females. These were correlated with gender- and wealth- inequality as well, and the reversal of direction for sex differences primarily reflected cultural factors. In Western nations with more egalitarian conditions, there was less pressure to pursue a STEM-based career for girls. But in countries with relatively low gender equality, pursuing a STEM-career means economic independence for women, and thus there is increased societal pressure to excel in such fields. Although a replication with a subsequent PISA wave failed to
support such conclusions and claimed that Reilly (2012) might represent a Type I error (Stoet & Geary, 2015), a subsequent study Stoet, Bailey, Moore and Geary (2016) did observe that countries with higher levels of gender equality showed larger sex differences in mathematics – an apparent paradox that did not cite prior research finding this effect (Guiso et al., 2008; Reilly, 2012). Nonetheless it is taken as a replication, even if not explicitly acknowledged.

Helgeson (2017) had argued that with the further passage of time since Hyde and Linn’s meta-analysis on verbal ability, there was a need for a stronger evidence base for testing such hypotheses. Helgeson reviewed the cross-cultural evidence presented in Chapter 6 and published as Reilly (2012), concluding that it showed sex differences remain robust for reading ability. Indeed, when reviewing this study Hyde (2014) acknowledged that the universal pattern of higher female performance in reading in all countries was “difficult to reconcile” (p. 382) with her earlier conclusions. That concession made me question whether sex differences might be present with other areas of verbal and language ability (leading to the analysis presented in Chapter 5). Miller and Halpern (2013b) also reviewed the cross-cultural analysis of PISA data presented in Chapter 6, devoting extended coverage of it and related studies, whereby they refined their psychobiosocial model of sex differences to include the contribution of macro-level cultural factors.

A recurring theme throughout this body of research is the words of sex difference researcher Professor Diane F. Halpern, who has remarked that in the field of sex differences “what you find depends on where you look” (Halpern, 1989, 2014a). Any investigator brings to bear their own ideological biases (alpha bias maximises, beta bias minimises), but Halpern has argued that ignoring the data, or neglecting to fully investigate it, does not advance the field or help reduce actual sex differences in
educational outcomes. Too often the focus of researchers has been in areas of male advantage (such as visual-spatial ability and quantitative reasoning), so it would have been remiss not to also investigate areas of potential female strengths such as reading and writing. The debate over the nature of sex differences will continue, but it is satisfying to have contributed to developing a stronger evidence base for evaluating such claims.

11.2 Contribution of sex-role identification to cognitive performance

The existence of sex differences in specific cognitive abilities has been examined, researched, and debated since the beginnings of psychometrics and measurement of human intelligence (Eagly, 1995; Maccoby & Jacklin, 1974). Yet despite over a century of psychological research, the question of their origins seemed an intractable enigma that has defied our best efforts to solve. The field had progressed from purely biological explanations (sexual dimorphism in brain structures, genetic/evolutionary contributions, and then endogenous sex hormones) to largely psychosocial explanations (sex differences in early socialisation experiences, differential treatment by parents and teachers, gender stereotypes). Archer (1996) has called these “origin theories”, a term that has since been adopted by other authors (e.g. Eagly & Wood, 1999). While these in isolation did not provide a satisfying explanation (or even explain a large portion of variance in the gender gap), pioneers like Halpern and Eagly brought the field towards embracing a biopsychosocial model of sex differences. But quite rightly, critics such as Hyde argue that the overlap between males and females is substantial, and rife with exceptions – males who perform poorly on visual-spatial/quantitative tasks, females who perform poorly on verbal and language tasks. How ought we explain these common exceptions to general rules about sex differences?
The sex-mediation hypothesis proposed by Nash (1979) had generated an initial burst of research interest, with enough studies conducted that Signorella and Jamison (1986) performed a meta-analysis to test the robustness of the effect for visual-spatial reasoning. This effect was replicated a decade later by Hamilton (1995) for other aspects of visual-spatial ability, including spatial visualization (GEFT). However few studies tested the second tranche of Nash’s theory, which was that femininity was associated with the cultivation of language and verbal abilities. Those studies that did (e.g. Ritter, 2004) were often hampered by serious methodological limitations. These included insufficient samples sizes for statistical power, employing only a single type of verbal measure, and a failure to calculate effect sizes of the comparison between those high/low in femininity in line with Nash’s hypothesis. Fortunately, such calculations can easily be performed from descriptive statistics. Subsequent calculation of effect sizes from Ritter’s reported descriptive statistics found that in the female sample, feminine ($d = -.77$) and androgynous ($d = -.33$) participants scored significantly higher than masculine ones. The effect size in the study reported by Ritter was compelling enough that I believed it merited providing a comprehensive test of Nash’ sex-role mediation hypothesis with verbal ability in a modern sample.

Before embarking on such an endeavour though, I wanted confidence (if only for peace of mind) that at least the visual-spatial aspect of the hypothesis would hold up to the passage of time. Chapter 7 and published as Reilly and Neumann (2013) offered ‘proof of concept’ that the effect was robust with the passage of time by conducting a new meta-analysis for mental rotation tasks. Surprisingly, quite a number of studies had measures of sex-role identification and mental rotation performance as part of a battery of neuropsychological tasks, but never examined the correlation as they were unaware of Nash’s hypothesis. A number of authors kindly reanalysed their data to provide these
correlations, leading to a much broader pool of studies than those explicitly testing the sex-role mediation hypothesis. With increased data, the effect seemed robust in both males and females. The review went further as well, by outlining the link between visual-spatial development and quantitative reasoning that more recent research had identified. It also offered a stronger rationale for further research beyond just an intellectual interest. Given the underrepresentation of women in STEM-fields other than medicine and psychology (National Science Foundation, 2017), and the prominence that encouraging visual-spatial reasoning was receiving in addressing this issue (Uttal, Miller, & Newcombe, 2013), further understanding of factors contributing to visual-spatial development has important practical implications.

Chapter 8, and subsequently published as Reilly, Neumann and Andrews (2016) aimed to provide a “thorough” test of the sex-role mediation hypothesis by having adequate statistical power to detect an effect. The experiment was designed to offer an array of visual-spatial and verbal language measures (in part, because a single measure could be easily dismissed as Type I error and chance, but also to examine the robustness of the effect and if it would generalise across types of tasks). As other researchers have noted, visual-spatial and verbal abilities are not unitary constructs (Halpern, 2011), and generalisability across tasks is vital if we are trying to infer an association with latent ability. Regardless of the experimental outcome, it was important that the second tranche of Nash’s hypothesis would had been adequately tested with a modern sample – if only to settle the question conclusively. For the domain of visual-spatial reasoning, the previously reported association between masculine sex-roles and mental-rotation was replicated (Reilly & Neumann, 2013). This also generalised to other types of visual-spatial tasks: there was a significant association for spatial visualization (GEFT) and spatial perception (Piaget WLT) tasks in both male and female participants,
consistent with the earlier meta-analysis by Signorella and Jamison (1986). Further, an
association was also found between feminine sex-role identification and all three verbal
and language tasks (verbal fluency, synonym generation, and DAT language usage). In
reviewing the study, Petersen (2018) argued that a sex-role mediation effect may help
explain the apparent sex difference in reading and writing outcomes through restricting
or promoting activities that provide additional training opportunities for reading and
writing skills, as well as cultivating verbal fluency.

Having demonstrated support for a sex-role mediation model for both verbal and
language ability, and for visual-spatial ability, it remains the task of future research to
elucidate the underlying mechanisms by which the sex-role mediation effect is realised.
One such mechanism for which there exists a robust body of research is that sex
differences in visual-spatial development are the result of differential levels of practice
and training between males and females, arising from the sex-typing of many leisure
activities and interests (e.g., model-making, sports, and computer games like Tetris,
Minecraft, etc.). As reviewed in Chapter 3, additional opportunities for spatial
development and enrichment lead to a substantial increase in visual-spatial abilities for
both males and females. While there is evidence from retrospective recollections in
adults of an association between visual-spatial ability and childhood spatial experiences
(Signorella, Jamison, & Krupa, 1989), it is likely that there are multiple mechanisms
involved such as situational factors of the testing environment and sex-role conformity
pressures (see Section 11.3). One would hypothesise that a similar effect of an
association between feminine sex-role identification and verbal and literary facilitating
experiences might also be found, but only a limited number of studies have tested this
mechanism directly. It is however consistent with McGeown et al., (2011) who found
that in children, a feminine sex-role identity was a better predictor of reading motivation
and reading ability than biological sex. More research is needed to see if this extends to other verbal and language tasks in children, including writing and vocabulary.

11.3 Contribution of situational factors to cognitive sex differences

While additional training opportunities may be one mechanism underlying the sex-role mediation effect, another equally plausible possibility is that the effects observed in Reilly et al. (2016) were not solely the result of differences in latent ability but also the product of situational factors of the testing situation itself. There were additional mechanisms outlined in Nash’s original theory concerning the perceived sex-typing of cognitive tasks (as either a masculine or a feminine one and hence requiring those traits stereotypically associated with either gender), and the knowledge of gender stereotypes about intellectual abilities. Might these situational factors better explain these observed results rather than actual ability?

This is, of course, not a new question – it is one that has been in the minds of sex difference researchers in other domains for some time, given the test/grades discrepancy for quantitative reasoning (see Section 2.2.3). Else-Quest et al. (2013) have argued that psychological traits (such as attitude, self-efficacy beliefs, self-concept) might also contribute to cognitive performance on standardised tests, as well as the decision to pursue, (or to not) STEM careers. Classical test theory (CTT) holds that performance on a test (raw score) is a function of true ability and measurement error. We often consider the fluctuations in psychological states during a testing session to be a source of transient error (Chmielewski & Watson, 2009; Schmidt, 2003; R. L. Thorndike, 1951); all things being equal systematic bias from these fluctuations will cancel out, and that the only challenge it poses for the purposes of research is increased “noise” and a challenge to statistical power. Measurement error does become problematic though at the individual level when high-stakes testing is used such as standardised tests of
educational achievement (e.g., entry tests of placement and admission to university or college). A common example of this is test-anxiety, whereby internal psychological states and excess physiological arousal have a deleterious effect on cognitive performance. However, there is a more insidious type of state effect that can impact measurement precision. This is when mood and state become a source of systematic error, in that they affect one demographic group disproportionately through mechanisms such as stereotype threat (Pennington, Heim, Levy, & Larkin, 2016; Spencer, Steele, & Quinn, 1999; Steele, 1997). So beyond testing whether a sex-role mediation effect was still present after controlling for state effects, an additional goal of Chapter 9 was to test whether situational factors (such as the perceived sex-typing of a cognitive task) could exert an influence on performance - even in a low-stakes testing environment where the participant was reassured that there was no reward or penalty for performance and data collection was anonymous.

The results of the study reported in Chapter 9 replicated earlier findings by showing that the way test content is perceived can substantially increase or decrease performance. When test content was portrayed as being a measure of empathy and perspective taking, women scored substantially higher \((d = .91)\) on the spatial visualization task than when it was portrayed as being visual-spatial in nature – despite sitting the same test content, under the same timing conditions, in the same room. This replicates the original study by Brosnan (1998) and a more recent study by Massa, Mayer and Bohon (2005) in a college-aged sample of women. Support for the sex-role mediation hypothesis was also found, with masculine and androgynous (high masculinity) women scoring significantly higher than feminine and undifferentiated (low masculinity) women. Concurrent support suggests that while situational perceptions of task content do exert an effect, there is still a meaningful contribution of
sex-role identification that reflects latent ability. Thus, sex-role identification can help to explain within-sex variability as an individual differences factor, as well as being a driving force behind between-sex variability.

While previous studies (Brosnan, 1998) had found an effect of task-labelling on female subjects performance on the GEFT spatial visualization task, there was the possibility that the effect would not replicate with a contemporary sample given changes in sex-roles and gender stereotypes. Further the Brosnan study had recruited primary school students, and the task-labelling of the GEFT as a measure of empathy and perspective taking may have been implausible with an adult sample. This is why the study included a second component on explicit priming of knowledge of gender stereotypes in the event that task labelling was too subtle. Contrary to expectations, the effect size difference between gender priming and control groups was somewhat lower for the mental rotation task ($d = .54$) than for spatial visualization, but the tasks are not directly comparable because mental rotation reflects different aspects of spatial ability and is also a more complex task. Again there was also a replication of the sex-role mediation effect for mental rotation observed in Reilly et al., (2016) in the sample of women, mounting stronger evidence for actual differences in ability.

Additionally, the study provides a cross-replication of the sex-role mediation effect for verbal and language abilities, at least in a sample of women. For the verbal fluency task, the high femininity groups generated significantly more words than the low femininity groups ($d = .47$). However, we did not see the predicted stereotype lift on the verbal fluency task – it may be easier to diminish performance through negative stereotypes than it is to raise performance through positive stereotypes (Walton & Cohen, 2003). It is a hypothesis seldom explored (e.g., Keller, 2007), and a direction future research might pursue is whether a similar stereotype threat exists for adolescent
and adult males who are completing verbal and language tasks. Two studies have found small stereotype threat manipulation effects for reading, but only in extremely young children (Hartley & Sutton, 2013; Pansu et al., 2016). The stereotype that men are less proficient in verbal and language tasks is widely held in adult samples (Halpern, Straight, & Stephenson, 2011; Swim, 1994), but to my knowledge no study has yet convincingly demonstrated stereotype threat for adult males on language tasks. It is an untested question that, had time permitted, I would have liked to address in a further study.

Another contribution of this study is to show that performance on cognitive tasks (even in low-stakes testing conditions) can be easily manipulated by situational factors such as the way the test is perceived or explicit priming of gender stereotypes. In that regard it has broader relevance to the question of how much of the gender gap in experimental studies and standardised tests can be explained by situational factors of the testing environment that affect women disproportionately to men, especially with high-stakes tests such as tertiary entry exams (Leiner, Scherndl, & Ortner, in press). It also helped strengthen my conviction that we need to move beyond simply studying sex differences on actual cognitive abilities to also consider what psychological trait and state variables explain individual differences in performance.

11.4 Contribution of sex role identification to self-estimated intelligence

Chapter 10 set out to explore a fundamentally different research question to that examined in the thesis so far, asking not about actual differences in intellectual ability but rather the way an individual perceives themselves to be different, especially relative to their peers. The power of expectations was illustrated in Rosenthal and Jacobson’s (1968) classic ‘Pygmalion in the Classroom’ experiment, where experimentally manipulated teacher expectations of the intelligence of randomly selected students
SEX AND SEX ROLE DIFFERENCES IN COGNITIVE ABILITIES

(identified as ‘gifted’) translated into higher psychometrically-measured IQ growth for those students over the course of several years. Most of the major reviews of sex difference research have focused primarily on actual ability (do differences exist, if so how large are they), but neglected to consider the way in which a person sees themselves – either as intellectually capable and on par or somewhat above their peers in intellectual ability, or instead as lacking the same intellectual prowess of those around them. What might that do, over the course of a child’s intellectual and educational development to their sense of intellectual self-competency, self-efficacy beliefs, and their patterns of academic interests? Research into academic motivation shows that perceptions of competence (Stipek & Gralinski, 1991, 1996), as well as beliefs about the rigidity or malleability of academic success, have important implications for how individuals progress in their education, especially in sex-typed intellectual domains such as STEM (Wang, Eccles, & Kenny, 2013).

Though sex differences in self-estimated intelligence (SEI) are frequently reported (Szymanowicz & Furnham, 2011), relatively few studies have attempted to examine their developmental trajectory. Some studies have investigated whether there are sex differences in the “intellectual self-concept” of children, and found either that very young girls and boys share similar self-concepts of their intellectual ability, or that girls start with slightly higher self-concepts (Aaron et al., 2005; Eccles, Wigfield, Harold, & Blumenfeld, 1993). Yet as early as fifth grade we can see demonstrable sex differences in intellectual self-concept, with boys seeing themselves as brighter (Gold, Brush, & Sprotzer, 1980; Marsh, 1989). By the time these children reach high school we see the robust sex difference in self-estimated intelligence (Steinmayr & Spinath, 2009), growing still larger in adulthood with college-aged samples. Intellectual self-concept and self-estimates of ability in academic domains plays an important role in
shaping achievement-related decision making (Ackerman & Wolman, 2007; Eccles et al., 1993; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002), so understanding of the antecedents of sex differences in self-estimated intelligence is critical for developing educational and motivational interventions.

However, the bulk of literature attests to the robustness of the effect (across types of samples, ages, socioeconomic groups, and cultures), but not its developmental antecedents. While a full investigation into this phenomenon would be best served by a longitudinal study (before sex differences in intellectual self-concept emerge) some information can still be gleaned by looking at the developmental end-point (young adults) to explore two hypotheses – firstly that the self-estimated intelligence might be explained by low self-esteem (which past research suggests is more prevalent in females), and secondly that it might be explained by patterns of sex-role identification.

Following the established protocol used by Furnham and Rawles (1995), participants were asked to provide an estimate of their intelligence (expressed as an IQ score), and then estimate their ability for specific cognitive domains as per Gardener’s multiple intelligences. A point of difference in this study is that participants also completed the Cattell’s Culture Fair Intelligence Test (CCFIT) as a measurement of psychometric intelligence, followed by measures of self-esteem and sex-role identity (BSRI).

The inclusion of an objective measurement of intelligence allowed us to rule out two possible explanations for what has been termed by Furnham et al., (2001) as the male hubris/female humility (MHFH) effect. Firstly, we could rule out the possibility of actual sex differences in intelligence due to convenience sampling from a student subject pool as one possible explanation, which had been a limitation of most previous studies. Secondly, by examining the correlation between objective and self-estimated
intelligence we were able to rule out a gross discrepancy in the accuracy rate of self-assessments: if, for example, males held utterly unrealistic evaluations of their intellect (no association) but females presented more realistic impressions. The results showed that the correlation strength was comparable across men and women, but what differed was the *direction* of reporting. More women underestimated their intelligence whereas more males overestimated. Interestingly this effect was not universal in that some males also underestimated their intellectual capabilities. Thus, the phraseology chosen by Furnham et al. as male hubris may be inapt, and the picture is more nuanced than first assumed.

Significant sex-role differences were also observed for self-estimated intelligence, with high masculinity participants reporting significantly higher IQ estimates than low masculinity groups, regardless of biological sex. Consistent with past literature (Whitley, 1984; Whitley & Gridley, 1993), there were also sex and sex-role differences in self-esteem for our sample. I had hypothesised that this might be one mechanism resulting in lower self-estimations of intelligence in women. Regression analysis showed that both masculine sex-role identification and self-esteem contributed to self-estimated intelligence even *after* controlling for psychometric intelligence. Of these, masculine sex-role identification exerted a stronger force. Mediation analysis confirmed that masculine sex-role identification acted as a mediator between sex and self-estimated intelligence scores, but there was still a significant direct effect of sex as well (partially mediated).

Sex differences were also found for many, but not all, estimations of the different components of Gardener’s multiple intelligences. The strongest effects were found for logical-mathematical and spatial intelligences, followed by bodily-kinaesthetic, naturalistic and existential intelligence. Significant differences were not
found for verbal intelligence (traditionally regarded as feminine), musical intelligence, and the two emotional intelligences (again, traditionally regarded as feminine). Thus the pattern of results seemed to be driven more by commonly regarded gender stereotypes (Swim, 1994), or what are lay intuitions about the direction of empirically observed cognitive sex differences (Halpern, Straight, et al., 2011). There was stronger evidence though for sex-role differences in specific multiple intelligences, with larger effect sizes and in more domains.

One question that remains somewhat unresolved in my mind is the extent to which sex differences in self-estimated intelligence are genuinely held, or whether they are the product of a gender-specific social desirability bias to either boast (in the case of males) or downplay (in the case of females) intelligence. The moderately-sized bivariate correlation between self-estimated intelligence, general and academic self-esteem offers some evidence that these are deeply held though. However, a follow-up study might consider including some measurement of what has been termed in the sex-role literature as felt pressure to conform to sex-role expectations (and in this case either enhance or to downplay personal intelligence). Some studies of mate preference have found that across cultures women hold a preference for partners that are smarter (Buss & Barnes, 1986; Shackelford, Schmitt, & Buss, 2005), but that men on average do not (Park, Young, Eastwick, Troisi, & Streamer, 2016). At a cultural level, a sex-specific pattern of continual self-enhancement/self-deprecation of intellectual ability may later transfer to genuinely held beliefs.

Finally, the study has identified sex/sex-role differences in self-esteem, and in particular academic self-esteem, as important targets for further study in understanding the factors contributing to sex differences in self-estimated intelligence. Given the abundant evidence that there is no support for actual sex differences in intelligence, and
that a cluster of males in our study also under-estimated their intelligence, educational interventions aimed at raising intellectual self-concept might be necessary for raising the educational aspirations of both genders.

11.5 Collective findings and implications for theory building

Although each of the individual studies comprising this thesis makes a relatively specific contribution, when viewed collectively they have broader implications for theory building and extending existing psychobiosocial models of sex differences in cognitive abilities. One of the most widely accepted models was that offered by Halpern (2004, 2011) as a biopsychosocial model emphasising the reciprocal relation between nature and nurture in the development of sex differences. This model was later endorsed in a consensus statement by many of the prominent sex/gender researchers in the field (Halpern et al., 2007). Though the model has its advantages (chiefly that it does not attribute weighting to any particular element, and thus has been better embraced by researchers who openly disagree about the relative emphasis of biological and social processes), it is light on details about the precise mechanisms by which sex differences develop. It also fails to explain why some types of cognitive tasks show higher performance in females but others show the reverse pattern of higher performance in males. These are themes that Halpern would return to in later works (Halpern et al., 2007; Halpern, Beninger, & Straight, 2011). Additionally, at the time it did not explicitly acknowledge the contributions of cultural factors (see Chapter 6), as an additional factor over and above one’s direct social environment. Subsequently, the model was refined by Miller and Halpern (2013b) to acknowledge culture as a pathway, which included extended coverage of this issue on pages 40 and 41 of their review including my own study published as Reilly (2012).
Another prominent model, and one that I have drawn from, is the biosocial model of sex differences proposed by Wood and Eagly (2012), which is based on Eagly’s (1987) social role theory. It highlights the social construction of gender through sex-roles, which serve to restrict and regulate thought and behaviour. At the heart of this model is an emphasis on the way that individuals self-regulate their behaviour to conform to sex-role stereotypes (which in turn are socially regulated through conformity pressures, rewards and sanctions). As such it is highly compatible with a sex-role mediation explanation for sex differences. Where my proposed model differs is that it omits the activational effect of sex hormones as a regulatory process, due to the relatively weak effect sizes and inconsistencies observed in the literature. While some researchers contend sex steroids such as testosterone and estrogen exert an activational effect on cognitive processes (e.g., Hampson, 2018), studies are frequently inconsistent and to some degree context-dependent to certain environmental cues (Halari et al., 2005; Hampson, 2018; Hausmann, Schoofs, Rosenthal, & Jordan, 2009; Puts et al., 2010), so this pathway has been intentionally omitted from the model at this time.

Collectively, the studies in this thesis can shed light on some of those mechanisms and can contribute to a more detailed psychobiosocial theory. One of the most powerful criticisms levelled at theories of sex differences (and their practical impact) comes from Hyde (2005, 2014) who has opined that within-sex variability is larger than between-sex variability, and that theorists would be better served explaining those individual differences rather than focusing on biological-based sex factors. Though forcefully conveyed in her work, it actually stems from an argument first made by Thorndike’s (1914) work into sex group differences and the role of individual differences factors in educational psychology. Any theory attempting to explain why sex differences emerge at the population level ought also to be able to explain sex-
related contributions to individual performance and within-sex variability. Thorndike was also the first to suggest that many of the causes of group differences were not due to innate biological differences, but rather to differential levels of training and practice arising from sex differences in activities and interests.

Nash (1979) had proposed a sex-role mediation theory of sex differences, but it was somewhat dated and focused primarily on social mechanisms. In proposing a revision of this theory, I have taken into consideration the two biopsychosocial models of sex differences described above, as well as more recent work on social and cultural factors.

Figure 11.1 presents a proposed biosociocultural model outlining such mechanisms. It emphasises how, initially very small, biological contributions as identified in the literature make a contribution to early brain development as relatively distal factors. The biological contributions include evolutionary pressures (see Section 2.3.1.3) manifesting as genetic predispositions differentiating the sexes, as well as the contribution of prenatal hormonal exposure. Collectively they exert a weak and indirect force on brain development as well as contributing to later psychosocial behaviour. More proximal is the contribution of early socialisation experiences which typically differ between boys and girls, but which are also subject to wide individual differences. Contributing to the acquisition of sex-role identification also are sociocultural factors, such as gender segregation in the division of occupational and family roles in society, the propagation of gender stereotypes (explicit and implicit), and the level of gender inequality in the society in which a child is raised. The relative contribution of biological, social and cultural factors is idiosyncratic to the individual resulting in individual differences in acquisition of masculine and feminine sex-role identification, but with enough commonalities differentiating the sexes that there are observable group
Figure 11.1 Biosociocultural model of sex-role identification acquisition, recognising the contribution of biological factors, early socialisation experiences arising from differential treatment of boys and girls, as well as cultural factors such as stratification in the roles of men and women in society, propagation of gender stereotypes and sex-typing of activities/intellectual interests.
differences between males and females as well in society. Thus the effect of sex (both as a biological factor, and a social category) and culture are mediated through sex-role identification. At earlier stages of development, environmental factors (such as availability of sex-typed toys, the type and quality of amount of communication and parental facilitation of verbal skills) may hold greater influence as parents, caregivers and teachers exert tight control over the types of experiences available to a child. But as children grow in autonomy, they start to gain greater control over their environment and begin to self-select activities and interests based on their personality and interests (niche-picking). Upon reaching adolescence - when conformity pressures increase – their own sex-role identification and personality traits can manifests as either a broad or narrow sex-typed repertoire of academic interests and leisure pursuits. Some elements are under the control of the child, and some elements (such as interactions with parents and teachers) reflect cultural values and gender-norms.

Figure 11.2 illustrates a sex-role mediation theory for the sex-typing and acquisition of cognitive skills. Like its predecessor, the model highlights the association between masculinity and development of visual-spatial ability and outlines a causal mechanism through differential exposure to spatial experiences and training, as well as sex-role conformity pressures. Visual spatial development is important not just as a skill in itself, but also because it lays down a foundation for the development of quantitative reasoning skills in mathematics and science. At the beginning of this course of research, the association between visual-spatial ability and quantitative reasoning skills was largely correlational in nature (e.g. Wai, Lubinski, & Benbow, 2009). However, recently published studies of educational interventions providing spatial training have found it delivers increased mathematics and science self-efficacy and more importantly, transfer
**Figure 11.2** – Sex-role mediation theory of cognitive development.
effects in the form of improved grades in college and university students (Miller & Halpern, 2013a; Sorby, Veurink, & Streiner, 2018). One study has also demonstrated similar outcomes as early as elementary school for mathematics (Cheng & Mix, 2014). As evidence for the pathway between development of visual-spatial ability and quantitative reasoning now stands on surer footing, and for this reason has been incorporated into the model.

Chapter 8 demonstrated support for a sex-role mediation effect with visual-spatial ability, with the evidence from Chapter 9 supporting both a difference in latent ability (arising from differential training from exposure to spatial experiences) as well as an effect of perceived sex-typing of spatial tasks. Additionally, the revised model in Figure 11.2 also identifies a new pathway that was not present in Nash’s original theory, which is the association between masculinity and self-estimated intelligence. This aspect of the model is based on the findings outlined in Chapter 10. This outcome is important, because of the role that perceptions of intellectuality play in voluntary course selection of more challenging academic content such as STEM subjects, as well as buffering against negative cultural gender stereotypes (outlined in Eccles’s (2007, 2013) expectancy-value model of academic achievement related choices).

Also presented in Figure 11.2 is the association between femininity and the cultivation of verbal and language abilities. It was originally postulated by Nash, but only a handful of studies had investigated this tranche of the theory. Chapters 8 and 9 investigated sex-role differences in several types of verbal fluency tasks, as well as grammar and language usage, finding support for the model. However verbal abilities are not a unitary construct, and incorporate a diverse range of tasks (see Section 2.2.1), and observance of a sex-role mediation effect with other tasks including reading and writing would further support the model.
To illustrate the utility of this revised model for explaining both between- and within-sex differences, I will give two hypothetical cases (one male, one female) to show how it might work in practice. Consider the case of James, a young boy of above average intellect who is born into a family with traditional beliefs about sex-roles and endorsement of gender stereotypes. In early childhood, James is provided with a variety of traditionally masculine toys and games – from cars and trucks that illustrate force and motion, to construction blocks that cultivate visual-spatial development. James’s mother talks with him often, but not to the same extent as his sisters about the same topics: there is more talk about practical things and activities and less about emotional feelings that support the scaffolding of social development and verbal aptitude (which given his moderately high IQ, he has the potential to excel in). Upon entering school, James finds he is called upon less to answer questions in language arts classes but more often in mathematics and science classes (his teachers hold high expectations for him, as he demonstrated an initial aptitude…. but there may be other similarly talented girls who have been overlooked because of gender stereotypes). Like many of his male peers he struggles with reading – it doesn’t come naturally to him, and his friends see it as a ‘girlie task’.

Already cultural and environmental factors are exerting an effect on his intellectual development. His father takes him to museums though and encourages him to learn coding, and slowly James finds himself acquiring an interest in STEM. In fact, it is something that he enjoys, and begins to self-select activities that suit his interests. Being strongly masculine sex-typed now though, James holds very little interest in literature; reading is viewed as necessary for school but not a source of intellectual stimulation and pleasure. Consequentially, James reads less often than does the typical child his age. As sex-role conformity pressures increase as he enters high school, he
holds expectations that he will probably pursue a career that draws on his strengths in science. He also has great confidence in his intellectual ability – his parents think James smart and tell him so all the time, and James feels he is much brighter than the typical student. He does well in standardized tests, as he goes into them with confidence.

The totality of his experiences up to that point have shaped both his personality and interests, magnifying his potential in some areas such as mathematics and science but blunting it in others such as verbal and language abilities. There were opportunities for his trajectory to diverge though: for example, if he’d been provided with toys and games that encouraged language development, if his parents had given him comic books to scaffold Jame’s early literary development followed by sci-fi or adventure novels, which might buffer against the gender-conformity pressures he would later experience in adolescence – well, James might equally have become a great writer and perhaps combined his love of STEM into a career of science journalism. Ultimately his strong masculine sex-typing might have limited his choices, but a greater variety of experiences might have tempered this.

Consider another hypothetical example, a girl named Sarah born to a different family but of equal innate intellectual potential as James. Sarah’s parents are both middle-class workers who realise the value in attaining an education, regardless of a child’s gender. Both her mother and father encourage her verbal communication skills which manifest slightly earlier than Jame’s in line with developmental effects. This gives Sarah a headstart with vocabulary development and Sarah’s parents both read to her nightly, scaffolding her emerging literacy skills. Together Sarah’s parents make a conscious effort to provide her with a range of educational toys and games, even ones that are traditionally masculine. Even though the toyshop they buy from is highly gender-segregated into separate boys and girls sections, Sarah’s father will often take
her shopping for Lego, especially the new ‘Friends’ line which holds a broader appeal. Tentatively, he also introduces her to age-appropriate transforming robots, and then scales them up in difficulty level as she grows. These encourage her visual-spatial development as well as cultivating a love for robotics and technology (she acquires the attitude that robots are ‘cool’). Although often seen as a masculine field, Sarah’s father encourages in her a love of science just as Jame’s did. She’s also exposed to a broader repertoire of experiences than other girls her age, encouraging a less rigid mindset about stereotypically masculine/feminine fields. She reads avidly, and like many girls her age enjoys writing – her teachers actively encourage this. She struggles with mathematics though, and feels that ‘it just comes naturally to boys’.

Sarah’s mother buys her a book on mathematics – not on how to do mathematics but on the important contributions made by women to the field throughout history. She falls in love with the tale of Countess Ada Lovelace, an exceptional mathematician who was the world’s first computer programmer. While mathematics in primary school is still a challenge, Sarah tells herself “if Ada could do maths, then I can too”. In high school her advanced visual-spatial skills relative to her female peers allow her to really come into her own (within-sex variability), facilitating the more advanced mathematics topics like geometry and trigonometry that make it possible to branch into studying physics and later chemistry. Compared to other girls her age, Sarah holds a greater interest in science, and is more open to the possibility of exploring further studies in a STEM field.

Sarah could become many things – a writer, an artist, a scientist or a doctor. She has attained a healthy blend of masculine and feminine personality traits (androgyny), and with that the consequence of behavioural flexibility in interests and talents. Her innate intellectual potential in STEM was not curtailed like so many of her female
peers, but she also has strong verbal and language talents that will compliment whatever career choice she chooses. She also has greater confidence - though still holds doubts - in her intellectual potential than many women her age (with a higher self-estimated intelligence in line with her IQ). Again, her trajectory could have been very different growing up in a more restrictive social environment, or if her own sex-typing had led her down different paths.

### 11.6 Directions for future research and limitations

While the literature and experimental studies presented in this thesis support the proposed sex-role mediation model of cognitive abilities, a chief limitation is that the experimental studies examined cognitive abilities at the developmental end-point (young adulthood). Definitive evidence of a mediation effect would require either cross-sectional or longitudinal studies to establish the effect of sex-role identity in younger students. Given that the meta-analyses presented in Chapters 4, 5 and 6 show that sex differences in educational outcomes have not yet been eliminated in modern samples, further research into their antecedents is sorely needed, and the optimal developmental stage would be before the gender gap in educational outcomes widens after puberty. However, in the case of self-estimated intelligence and what developmental psychologists term intellectual self-concept, this begins even earlier with differentiation between males and females appearing as early as fifth grade (Gold et al., 1980; Marsh, 1989). The psychological mechanisms behind this particular timing are not yet clear, and investigating whether nascent sex-role identities and endorsement of gender stereotypes (explicit and implicit) are responsible with younger children would be a useful research goal, as well as any intermediary constructs/processes.

Another important research goal is to further test the underlying mechanisms behind sex differences in verbal and language abilities. The proposed sex-role mediation
model outlines two primary processes that are responsible: namely differential levels of exposure to verbal and language tasks, as well as sex-typing. Meta-analyses of parent-child communication studies have shown that mothers engage in more talk with daughters than sons (Leaper, Anderson, & Sanders, 1998), girls are slightly more talkative than boys (Leaper & Smith, 2004), and that girls show much higher reading motivation in primary school (Marinak & Gambrell, 2010) and greater time spent reading for leisure than boys in high school which is positively correlated with reading achievement (Durik, Vida, & Eccles, 2006). Thus there is evidence for differential levels of practice and training between the sexes as one mechanism underlying sex differences, but further evidence is needed to document a sex-role mediation effect in children and adolescents. At present, few studies have investigated associations between sex-role identification, perceived sex-typing of language tasks, and performance on reading and writing tasks with younger age-groups. McGeown, Goodwin, Henderson and Wright (2011) found that feminine sex-role identification was a better predictor of reading motivation than biological sex, while Pajares and Valiante (2001) found the same pattern in a sample of primary school students for writing ability, motivation and self-efficacy. Importantly none of the studies found a negative association with masculinity. But replication of this effect with other samples would be desirable, as well as for other types of verbal and language abilities. This highlights sex-role identification as a useful construct in understanding individual differences in performance.

Another as yet unfinished task is to further investigate cross-cultural contributions to sex differences in educational outcomes. Chapter 6 reported as Reilly (2012) examined the impact of national levels of gender inequality on reading, mathematics and science, and a subsequent conference paper by Reilly (2015) replicated those findings for reading with subsequent PISA waves. However, further analysis
needs to be conducted to test whether the association with gender inequality replicates for mathematics and science achievement. There may be other, as yet unexamined cultural factors that explain the high degree of heterogeneity in effect sizes globally for mathematics and science, and why some nations show higher male performance but others show a reversal with higher female performance. There is tentative evidence that stark inequality leads to greater pressure on females to seek STEM careers, especially in non-Western nations. As PISA and TIMSS continue to integrate addition partner nations from the developing world, it will provide a good testing ground for these research questions.

11.7 Practical Implications for Childhood Education

The research contained herein raises some important practical implications for educational practice and the development of potential interventions. Firstly, it shows that substantial sex differences remain in verbal and language abilities, and that educators and researchers may have underestimated just how large a gap exists for writing tasks. This has important implications for boys’ preparedness to pursue tertiary education, and highlights the need for greater concentration on writing practice in the curriculum. Just as a greater focus on science and mathematics is important for encouraging girls in these domains, so too is a focus for literacy, the mechanics of grammar, and practice on writing tasks important for equality of outcomes with boys.

Secondly, this body of research demonstrates that males and females are not homogenous groups, and that sex-role identification (and the behavioural consequences thereof, in terms of differential training and conformity pressures) may explain a meaningful portion of the within-sex variability noted by earlier researchers such as Thorndike (1914) and Hyde (2005). For children who are highly sex-typed and have a narrowly constrained range of experiences and interests, there may be some merit in
gently attempting to broaden their experiences. Examples of which would include introducing stereotypically masculine spatial promoting toys and games, encouraging more verbal communication and social play, and scaffolding early literacy skills. Parents and caregivers are the initial gateway through which such experiences are provided, but as children grow in autonomy they begin to self-select; if they have previously been exposed to only a narrow range then their outlook may be similarly restricted and fall along stereotypically masculine/feminine lines.

Differential levels of practice through play and leisure activities is only one aspect of the sex-role mediation theory, however. Another important element is sex-role conformity pressures (which intensify in adolescence). Children acquire gender-stereotyped beliefs early, and they take cues from many sources – including parents, teachers, media and peers. Parental attitudes to reading and writing (or mathematics and science) convey important messages about the sex-typing of these pursuits, as does the way they are taught in school and which students are called on in class. Parental and teacher attitudes may be equally appropriate targets for interventions, as well as those of children. If children see either language or STEM as equally appropriate (indeed, expected) for both boys and girls, then this might influence the attitudes they bring to bear in adolescence and entering high school. While educational interventions often focus on increasing ability, equally important are student attitudes to the sex-typing of fields and showing them that they may be relevant to their future career interests.

Thirdly, the issue of sex differences in intellectual self-concept (as measured by self-estimated intelligence, and self-efficacy beliefs) really has not received enough research attention. Many – but not all – males significantly overestimate their intellectual abilities, while a substantial portion of females underestimate it. The reasons are not yet clear (e.g. differential levels of social desirability of intelligence between the
sexes, effects of child-gender on parental estimates of intelligence, cultural stereotypes), but our research suggests that sex differences in general self-esteem may play a substantial role. While there were group differences between males and females in our study, sex-role identification mediated this effect, with a masculine sex-role identification buffering against low self-esteem and leading to higher self-estimates of intelligence. This effect was present in both men and women. Males with low masculine identification are at particular risk, as are non-androgynous females. A holistic approach to reducing sex differences in educational outcomes should also target intellectual self-concept, as research has shown this strongly influences course selection in high school and career aspirations.

One issue that I feel important to address as well is the question of whether sex differences are an inevitable outcome of society, and whether attempts to influence childrens’ play experiences and sex-role beliefs constitute social engineering. Tackling the first question, cross-cultural research demonstrates that there is substantial variability in mathematics and science outcomes. In some nations, there are strong patterns of higher male performance, in others there are no significant sex differences, and in others substantially higher female performance is evident. The latter are often countries where gender and income inequality are substantial, and STEM represents a pathway to economic independence. None of these findings suggest sex differences in quantitative reasoning is inevitable. The picture is less optimistic for reading achievement. Substantial sex differences are present in all participating nations, but these is also substantial heterogeneity in their magnitude. We simply do not have sufficient data to make conclusions about cross-cultural patterns of writing, but it is likely to be the same. It may always be the case that some sex differences in verbal and language abilities will exist, but as a matter of equity, I believe that parents, educators
and policy-makers should make concerted efforts to mitigate against them. Not every child will become a wordsmith or avid reader, but we may be able to raise the literary skills of the typical boy with educational interventions that have broader targets (spelling, grammar, verbal fluency, and writing) than just reading alone. And it may better prepare them for an uncertain future in an age of automation, where reskilling and further education will be expected of them.

The second issue may be seen as more problematic, as it has at its core elements of public policy and ideology that are contentious. Does society have a right – or even, an obligation – to attempt to engineer a child’s masculine and feminine identification with the aim of minimising sex differences in educational outcomes? Society is best served by diversity, and a plurality of different perspectives, with no sex-role category being more worthy or legitimate than another. And given that sex-role identification is determined, to a large degree, by biological forces and personality, will attempts at ‘modification’ be futile anyway? Even in studies with non-human primates, toy selection preferences are often innately driven and devoid of cultural influence. But certain combinations of sex-role identification have behavioural consequences, such as restricted or broadened interests, leisure pursuits, and academic attitudes. Here, parents and educators can make a contribution to the reduction of gender stereotypes and sex-typing of intellectual pursuits, without seeking to change a child’s temperament. Just as sports are now recognised as being equally important for girls as for boys, encouraging verbal and language proficiency in boys and STEM skills in girls may be important mechanisms for mitigating sex differences in educational outcomes. And building self-confidence and encouraging realistic intellectual self-concepts will be beneficial for any child: masculine, feminine or androgynous.
11.8 Final Conclusions

Sex differences in specific cognitive abilities is an actively researched but contentious topic, which has important social, occupational and public policy implications for the ways in which children are educated and supported (both formally in schools and informally in the home). The meta-analyses presented in this thesis have addressed a long-standing research question about the magnitude of sex differences, extending those of earlier researchers. Legitimate concerns had been raised that sex differences may have been declining or even eliminated in modern samples, but to paraphrase Mark Twain rumours of their death have been greatly exaggerated.

The notion that there might be inequality of educational outcomes for males and females rankles at our sense of basic fairness. Yet designing educational interventions or changes to teaching pedagogy requires a clearer understanding of why sex differences develop (origin theories). This has been a seemingly intractable research question asked since the beginning of psychometrics and intelligence testing. The sex-role mediation explanation examined herein identifies and validates an individual differences factor for explaining cognitive performance – both group differences between males and females, as well as within-sex variability argued by Thorndike (1914) and Hyde (2005). Rather than biological sex exerting direct influences, it is an individual’s combination of masculinity and femininity that shapes the development of cognitive abilities through differential exposure to stereotypically masculine or feminine past-times and intellectual interests. It also highlights how other (non-intellectual) factors such as self-estimated intelligence may contribute to disparities in educational outcomes, and how these are also associated with sex-role identification. The information gleaned in the present research may prove useful for designing educational interventions (such as targeting perceptions of sex-typing, or addressing differential levels of training). This thesis may
be a modest small step forward in the context of the enormity of the sex difference
debate, but the road to eliminating sex differences in educational outcomes will be long
and winding. In this respect, this thesis may allow researchers to travel around the
corner of mere debate about the presence of sex differences to head down a straight to
better understand how sex differences appear.
References


Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence, 48*, 137-151. doi:10.1016/j.intell.2014.11.006


## Appendix A1 – SAT Mathematics Meta-Analysis

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean (S.D.)</th>
<th>Male Mean (S.D.)</th>
<th>Variance Ratio (VR)</th>
<th>Cohen’s d</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>527 (115)</td>
<td>492 (107)</td>
<td>1.16</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1997</td>
<td>530 (114)</td>
<td>494 (108)</td>
<td>1.11</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1998</td>
<td>531 (114)</td>
<td>496 (108)</td>
<td>1.11</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1999</td>
<td>533 (115)</td>
<td>495 (110)</td>
<td>1.09</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2000</td>
<td>533 (115)</td>
<td>498 (109)</td>
<td>1.11</td>
<td>0.31</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2001</td>
<td>533 (115)</td>
<td>498 (113)</td>
<td>1.04</td>
<td>0.31</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2002</td>
<td>534 (116)</td>
<td>500 (110)</td>
<td>1.11</td>
<td>0.30</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2003</td>
<td>537 (116)</td>
<td>503 (111)</td>
<td>1.09</td>
<td>0.30</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2004</td>
<td>537 (116)</td>
<td>501 (110)</td>
<td>1.11</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2005</td>
<td>538 (116)</td>
<td>504 (111)</td>
<td>1.09</td>
<td>0.30</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2006</td>
<td>536 (117)</td>
<td>502 (111)</td>
<td>1.11</td>
<td>0.30</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2007</td>
<td>533 (116)</td>
<td>499 (110)</td>
<td>1.11</td>
<td>0.30</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2008</td>
<td>531 (119)</td>
<td>500 (113)</td>
<td>1.11</td>
<td>0.27</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2009</td>
<td>534 (119)</td>
<td>499 (113)</td>
<td>1.11</td>
<td>0.28</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2010</td>
<td>531 (119)</td>
<td>499 (113)</td>
<td>1.11</td>
<td>0.27</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2011</td>
<td>530 (123)</td>
<td>499 (114)</td>
<td>1.16</td>
<td>0.26</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2012</td>
<td>527 (124)</td>
<td>496 (115)</td>
<td>1.16</td>
<td>0.26</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2013</td>
<td>524 (126)</td>
<td>494 (116)</td>
<td>1.18</td>
<td>0.25</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2014</td>
<td>527 (115)</td>
<td>492 (107)</td>
<td>1.16</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2015</td>
<td>530 (114)</td>
<td>494 (108)</td>
<td>1.11</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2016</td>
<td>531 (114)</td>
<td>496 (108)</td>
<td>1.11</td>
<td>0.32</td>
<td>&lt;.001 ***</td>
</tr>
</tbody>
</table>

Point estimate $d = +.30$ [95% CI = .29 to .30], with males scoring higher than females.

Datasource: The College Board Archived SAT Reports
### Appendix A2 – SAT Verbal Meta-Analysis

<table>
<thead>
<tr>
<th>Year</th>
<th>Male Mean</th>
<th>Male (S.D.)</th>
<th>Female Mean</th>
<th>Female (S.D.)</th>
<th>Variance Ratio (VR)</th>
<th>Cohen’s d</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>507</td>
<td>110</td>
<td>503</td>
<td>110</td>
<td>n/a</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1997</td>
<td>507</td>
<td>111</td>
<td>503</td>
<td>111</td>
<td>n/a</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1998</td>
<td>509</td>
<td>111</td>
<td>502</td>
<td>111</td>
<td>n/a</td>
<td>0.06</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1996</td>
<td>509</td>
<td>111</td>
<td>502</td>
<td>111</td>
<td>n/a</td>
<td>0.06</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1997</td>
<td>507</td>
<td>111</td>
<td>504</td>
<td>111</td>
<td>n/a</td>
<td>0.03</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1998</td>
<td>509</td>
<td>111</td>
<td>502</td>
<td>111</td>
<td>n/a</td>
<td>0.06</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>1999</td>
<td>507</td>
<td>111</td>
<td>502</td>
<td>111</td>
<td>n/a</td>
<td>0.05</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2000</td>
<td>512</td>
<td>111</td>
<td>503</td>
<td>111</td>
<td>n/a</td>
<td>0.08</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2001</td>
<td>512</td>
<td>112</td>
<td>504</td>
<td>112</td>
<td>n/a</td>
<td>0.07</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2002</td>
<td>513</td>
<td>112</td>
<td>505</td>
<td>112</td>
<td>n/a</td>
<td>0.07</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2003</td>
<td>505</td>
<td>114</td>
<td>502</td>
<td>111</td>
<td>n/a</td>
<td>0.03</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2004</td>
<td>503</td>
<td>114</td>
<td>500</td>
<td>111</td>
<td>1.05</td>
<td>0.03</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2005</td>
<td>502</td>
<td>114</td>
<td>499</td>
<td>110</td>
<td>1.05</td>
<td>0.03</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2006</td>
<td>502</td>
<td>114</td>
<td>497</td>
<td>110</td>
<td>1.07</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2007</td>
<td>504</td>
<td>114</td>
<td>498</td>
<td>111</td>
<td>1.07</td>
<td>0.05</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2008</td>
<td>500</td>
<td>116</td>
<td>495</td>
<td>112</td>
<td>1.05</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2009</td>
<td>498</td>
<td>116</td>
<td>493</td>
<td>112</td>
<td>1.07</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2010</td>
<td>499</td>
<td>117</td>
<td>494</td>
<td>112</td>
<td>1.07</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2011</td>
<td>499</td>
<td>118</td>
<td>495</td>
<td>113</td>
<td>1.09</td>
<td>0.03</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2012</td>
<td>497</td>
<td>119</td>
<td>493</td>
<td>113</td>
<td>1.11</td>
<td>0.03</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2013</td>
<td>495</td>
<td>120</td>
<td>493</td>
<td>114</td>
<td>1.11</td>
<td>0.02</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2014</td>
<td>507</td>
<td>110</td>
<td>503</td>
<td>110</td>
<td>1.11</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2015</td>
<td>507</td>
<td>111</td>
<td>503</td>
<td>111</td>
<td>1.05</td>
<td>0.04</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>2016</td>
<td>509</td>
<td>111</td>
<td>502</td>
<td>111</td>
<td>1.05</td>
<td>0.06</td>
<td>&lt;.001 ***</td>
</tr>
</tbody>
</table>

n/a indicates male/female standard deviation data not available. Effect size calculated from pooled SD reported in College Board Archived SAT reports.

Point estimate $d = +.045$ [95% CI = .038 to .053], with males scoring slightly higher than females. Note that this direction is contrary to the generally reported sex differences in verbal and language abilities (see Chapter 2, Section 2.2.1 and Chapters 5 and 6).
Appendix A3 – Academic Self-Esteem Measures

The following items were used as a measure of academic self-esteem, in Chapter 10. They were adapted from items in Johnson et al.’s (1983) Academic Self-Esteem subscale, and Bachman’s (1970) Self-Concept of Ability Scale (SCAS), with wording changed from high school to university. Several items were reverse-coded (indicated by an asterisk) to minimise acquiescence bias.

<table>
<thead>
<tr>
<th>Academic Self-Esteem</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I feel confident in my academic abilities</td>
<td>SA</td>
<td>A</td>
<td>D</td>
<td>SD</td>
</tr>
<tr>
<td>2. I am not doing as well at university as I would like to*</td>
<td>SA</td>
<td>A</td>
<td>D</td>
<td>SD</td>
</tr>
<tr>
<td>3. Coursework is fairly easy for me</td>
<td>SA</td>
<td>A</td>
<td>D</td>
<td>SD</td>
</tr>
<tr>
<td>4. I sometimes feel lost in lectures and reading textbooks*</td>
<td>SA</td>
<td>A</td>
<td>D</td>
<td>SD</td>
</tr>
<tr>
<td>4. Whenever I take a test I am afraid I will fail or do badly*</td>
<td>SA</td>
<td>A</td>
<td>D</td>
<td>SD</td>
</tr>
<tr>
<td>5. I feel confident in my ability to complete university</td>
<td>SA</td>
<td>A</td>
<td>D</td>
<td>SD</td>
</tr>
</tbody>
</table>

For comparability, the same response format as the Rosenberg Self-Esteem Scale that participants completed on the previous page of their booklet was employed.

Additionally, we administered the single-item Rosenberg Academic Self-Esteem Scale which asks respondents to compare themselves to other students enrolled in their degree.

1. How do you rate yourself in academic ability compared with those studying your degree? (CIRCLE ONE)

<table>
<thead>
<tr>
<th>Far below average</th>
<th>Slightly below average</th>
<th>About average</th>
<th>Slightly better</th>
<th>Far above average</th>
</tr>
</thead>
</table>